Sussurro - Detecção na Web de eventos auditáveis que representam riscos à saúde pública

Raphael Silva Fontes [1,2], **Methanias Colaço Júnior** [1,2,3], **Helder Prado** [2,3], **Ana Nely** [4], **Júlio Araújo** [4], **Jailton Paiva** [1,2], **Ricardo Valentim** [1,2]

[1]Laboratório de Inovação Tecnologica – Universidade Federal do Rio Grande do Norte
Hospital Universitário Onofre Lopes - HUOL
Av. Nilo Peçanha, 620 - Petrópolis Natal - RN, 59012-300

{raphael.fontes, methanias.colaco}@lais.huol.ufrn.br

{jailton.paiva, ricardo.valentim}@lais.huol.ufrn.br

[2]Núcleo Avançado de Inovação – Instituto Federal do Rio Grande do Norte
Natal/RN, Brasil

helder.prado@navi.ifrn.edu.br

[3]Programa de Pós-graduação em Computação – Universidade Federal de Sergipe

[4]Auditoria-Geral do Sistema Único de Saúde – Minsitério da Saúde
Brasília/DF, Brasil

{julioc.aracaju, ana.nely}@saude.gov.br

Abstract. Context*: Despite the advancement of technology, many processes, especially in the public sector, still require manual or non-intelligent searches to build knowledge. For audits by the Ministry of Health and the Unified Health System (SUS), the various sources of research to be explored cause delay and high costs.* ***Objective****: Evaluate preliminarily a tool, based on textual content retrieval, the Sussurro, to collect and present the various matters related to public health, which can guide audits.* ***Method****: Sussurro proof of concept.* ***Results****: Among more than 2 million material records, 56,053 were classified as evidence for health audit.* ***Conclusion****: Sussurro optimizes the selection of content needed to perform an audit, benefiting the investigation and fight against corruption in healthcare.*

Resumo. Contexto*: Apesar do avanço da tecnologia, muitos processos, especialmente no setor público, ainda necessitam de buscas manuais ou não inteligentes, para construção de conhecimento. Para as auditorias do Ministério da Saúde e do Sistema Único de Saúde (SUS), as diversas fontes de pesquisa a serem exploradas provocam demora e altos custos.* ***Objetivo****: Avaliar preliminarmente uma ferramenta, baseada em recuperação de conteúdo textual, o Sussurro, para coletar e apresentar as diversas matérias relacionadas à saúde pública, as quais podem nortear auditorias.* ***Método****: Prova de conceito do Sussurro.* ***Resultados****: Entre mais de 2 milhões de registros de matérias, 56.053 foram classificadas como indícios para auditoria em saúde.* ***Conclusão****: O Sussurro otimiza a seleção de conteúdo necessário para executar uma auditoria, beneficiando a investigação e o combate à corrupção na área de saúde.*

1. Introdução

A saúde é um direito fundamental do ser humano, previsto no Artigo 196 da Constituição Federal do Brasil, portanto, é dever do Estado prover as condições indispensáveis ao seu pleno exercício [Brasil 1988]. O Sistema Único de Saúde (SUS), criado a partir da promulgação desta mesma constituição e inspirado pelo National Health Service do Reino Unido, permite que os cidadãos tenham acesso a serviços de saúde, tais como atendimentos primários, emergências e procedimentos complexos [Lima 2019].

A tecnologia em saúde pode ser exemplificada como uma ferramenta fundamental para resolver os problemas em sistemas de saúde. Desde dezembro de 2005, as tecnologias em saúde englobam sistemas de informações, programas e protocolos de assistência que provêm cuidados com a saúde para a população [Brasil 2010]. Neste contexto, nos últimos anos, foram concebidas diversas melhorias para o funcionamento adequado do SUS, tais como a melhoria de alguns processos e a sua informatização, as quais fortalecem a transparência das ações e o acompanhamento dos recursos financeiros destinados à saúde [Brasil 2015].

Diante desta realidade, torna-se primordial a participação de órgãos que assumam a responsabilidade de controle e fiscalização, visando a melhoria da qualidade dos serviços de saúde prestados pelo SUS. As atividades de auditoria, que são realizadas no Ministério da Saúde, contribuem diretamente para a gestão e para a utilização adequada dos recursos, com a garantia do acesso e da qualidade na atenção à saúde [Costa 2021]. A autoridade de auditoria nacional responsável pelo SUS, com as atividades de planejamento, monitoramento, avaliação, regulação, vigilância em saúde e de outros órgãos integrantes do sistema de controle interno e externo é o Departamento Nacional de Auditoria do SUS, DENASUS, que teve seu nome atualizado para AudSUS, em 2023 [Brasil 1993] [Brasil 2023].

O processo seguido pela AudSUS é um conjunto de atividades executadas para garantir a harmonia e a lógica de seus preceitos, os quais procedem para que os trabalhos sejam realizados com segurança, qualidade e consistência. A equipe de auditoria deve conduzir as respectivas atividades e elaborar documentos que contenham as tarefas a serem executadas. Dentre as realizações exercidas nas tarefas, existem três fases: analítica, operativa ou in loco e relatório final [Brasil 2017].

A fase analítica corresponde ao planejamento da auditoria, para que esta seja adequadamente executada, dentro do prazo estabelecido. Sua finalidade é de preparar a equipe de auditoria para a fase operativa, proporcionando o desenvolvimento de uma compreensão mais acurada sobre o contexto das atividades posteriores. Para a construção do conhecimento necessário para o trabalho, deve ocorrer o levantamento de informações sobre o objetivo da auditoria. Este levantamento pode surgir a partir de diversas fontes, desde que permitam à equipe estabelecer uma visão geral sobre o objeto e seu contexto, tais como: auditorias realizadas anteriores pela AudSUS, sistemas de informação, relatórios do órgão responsável, sites da internet, bases de legislação e normas, atividades de controle realizadas por outros órgãos, artigos acadêmicos, informações da mídia e outras fontes relevantes [Brasil 2017].

Encarece que os auditores da AudSUS são responsáveis por fiscalizar todas as áreas do SUS e, além disso, também respondem às demandas internas e externas do Mi-

nistério da Saúde, tais como do Tribunal de Contas Federal, Ministérios Públicos Estaduais e Federal, Controladoria Geral da União, Controladorias Gerais dos Estados e Tribunais de Contas dos Estados. Com a quantidade de demandas e o atual efetivo de servidores, que conseguem auditar em média 100 processos por ano, o tempo necessário para zerar o passivo do Programa Farmácia Popular, por exemplo, ultrapassa 20 anos [Aquino 2022].

Neste contexto, este trabalho tem o objetivo de apresentar os resultados preliminares da ferramenta Sussurro, um classificador de notícias, baseado em Inteligência Artificial, para auxiliar os auditores na descoberta de informações oriundas de acontecimentos publicados na internet, as quais são capazes de priorizar investigações e contribuir com os processos de auditoria em saúde pública realizados pela AudSUS.

As demais seções deste artigo estão dispostas conforme: a seção 2 apresenta os trabalhos relacionados, na seção 3, estão os materiais e métodos, na seção 4, ocorre a demonstração dos resultados e as suas discussões, e, finalmente, na seção 5, estão contidas as considerações finais e trabalhos futuros.

2. Trabalhos relacionados

Apesar do grande potencial de pesquisa existente na utilização de notícias da internet para descoberta de indícios para auditoria e/ou fraudes no sistema público de saúde, utilizando processamento de linguagem natural, a exploração acadêmica ainda é escassa. Essa afirmação dar-se-á pelo fato de terem sido encontrados poucos trabalhos utilizando computação inteligente e notícias como objetos para identificação de fraudes e prioridades para auditorias públicas na saúde. Para que haja um embasamento sobre o estado da arte, estão referidas, a seguir, pesquisas com propostas que possuem alguma similaridade com o objetivo deste trabalho.

O trabalho de Simões e Costa [Costa 2021] utilizou a rede social Twitter para a descoberta de potenciais epidemias. Este trabalho apresenta contribuições relevantes na seção de análise, comparando as publicações e os acontecimentos durante a pandemia de COVID-19. Além disso, a pesquisa apresentou uma revisão sistemática da literatura, caracterizando o potencial da rede social digital para a descoberta de epidemias [Costa 2021]. A rotulação dos tweets e remoção de duplicidade foi manual e não foram apresentados modelos conceituais e de arquitetura do software. Além disso, também não há o detalhamento de como foi realizado o pré-processamento e o que foi desconsiderado nesta etapa, tal como: acentuação gráfica, números, emojis, caracteres especiais, dentre outras possibilidades.

O Weichselbraun [Weichselbraun 2020] apresentam uma solução baseada em Deep Learning, agregada às técnicas conhecidas de NLP (Natural Language Processing) para análise de sentimento, de palavras-chaves e de entidades nomeadas, com o objetivo de identificar potenciais riscos de corrupção. Este trabalho apresenta os resultados em dashboards e é capaz de identificar apenas as línguas inglesa e alemã, coletando e analisando as matérias publicadas nos principais meios de comunicação dos Estados Unidos da América, Áustria, Alemanha, Suíça e Reino Unido [Weichselbraun 2020]. A pesquisa não leva em consideração a ligação oportuna entre os casos publicados por órgãos governamentais e documentos internos dos seus clientes. Por fim, não há referência à possibilidade de rotulação de uma ou mais categorias para a notícia, bem como ao modo

de extração das notícias.

O trabalho de Diakopoulos [Diakopoulos 2020] apresenta uma arquitetura para orientação de publicação editorial, na qual as notícias são rotuladas como “interessante” ou “fraca”. Para validação do trabalho, o autor efetuou um estudo de caso no qual os participantes foram entrevistados para avaliação dos resultados esperados [Diakopoulos 2020]. Esta pesquisa não menciona em quais línguas é possível aplicar a ferramenta, bem como não demonstra se há alguma limitação quanto à categoria da notícia.

Por fim, o diferencial do trabalho aqui proposto é o acoplamento de diversas fontes e a classificação de indícios de auditoria, tornando-se um ”Google ++”da auditoria em saúde pública.

3. Materiais e métodos

Como proposta de avaliação inicial do Sussurro, foi elaborada uma prova de conceito, a qual, inicialmente, foi concebida a partir de entrevistas com os auditores da AudSUS, com o objetivo de elucidar o processo de captação do material utilizado durante a fase analítica da auditoria. Este material é colhido a partir da consulta de matérias existentes em: sites de notícias sobre saúde, relatórios técnicos, recomendações e portarias que são expedidos pelos órgãos de controle, diários oficiais, dentre outros. Nas entrevistas, também foram destacados os pontos-chave a serem extraídos de cada fonte, perfazendo o conhecimento necessário para execução dos trabalhos.

Em seguida, foi realizado um estudo exploratório com o objetivo de mapear como estão dispostas as matérias textuais presentes nas fontes, qual a rotina de publicação, periodicidade, quais as limitações e o que pode ser feito em casos de dados ausentes ou incompletos, tais como, por exemplo, quando não há a data de publicação ou o autor. Para os casos em que não há dados, foi utilizado o texto padrão “Não informado”. Com a finalização deste mapeamento, deu-se a construção de um modelo responsável por armazenar estes dados. Este modelo, de forma sintética, é construído de modo semi-estruturado, possuindo campos que são comuns entre as fontes, tais como data de publicação, título e sumário, bem como possibilitando que novas fontes e tipos de matérias com campos singulares sejam inseridos, sem que ocorra alteração na estrutura lógica do modelo.

Ato contínuo, ocorreram as etapas de coleta das matérias textuais, pré-processamento do conteúdo, mineração, classificação e a de disponibilização destes documentos para consulta por parte dos auditores. Essas etapas são detalhadas a seguir.

3.1. Coleta das matérias

Para o *pipeline* de coleta dos dados, é importante observar que cada fonte possui uma estrutura singular de publicação e apresentação das matérias. Por esse fato, utilizando Python com o Framework Django, banco de dados Postgres e OpenSearch para indexação dos resultados, foram construídos robôs especialistas, exclusivamente capazes de ler, entender e coletar os metadados.

Antes de ocorrer a coleta das matérias, é possível configurar e automatizar como os robôs devem se comportar na execução de suas atividades, sendo possível definir a busca por novas matérias ou o reprocessamento de publicações anteriores. Estes ajustes permitem, por exemplo, que haja repetições de buscas mais frequentes em fontes de

notícias que possuem publicações de matérias durante todo o dia, em detrimento às fontes que fazem uma publicação geral apenas no final do dia ou uma vez ao mês, como o caso de relatórios de auditoria e diários oficiais. Além disso, os robôs são capazes de entender quando há paginação, quando uma matéria já foi coletada, quando houve alteração, bem como onde inicia e termina o corpo textual da matéria, coletando somente o necessário e ignorando cabeçalhos, rodapés, menus extras e publicidades. Toda essa inteligência aplicada aos robôs permite que estes sejam capazes de capturar novas, algumas ou todas as matérias que houver na fonte.

Até o momento da escrita deste trabalho, os robôs especialistas adquiriram mais de 2 (dois) milhões de matérias textuais das fontes de pesquisa elencadas pelos auditores.

3.2. Pré-processamento

Após a coleta, essas matérias vão para o *pipeline* de pré-processamento, responsável por higienizar o texto e padronizar seus metadados no modelo estabelecido. Este processo é importante para garantir a qualidade e o armazenamento correto dessas informações no banco de dados e na ferramenta de indexação.

A atividade de higienização dos dados foi dividida em duas etapas: (1) a primeira é responsável por manter a estrutura padrão da notícia, contendo tabelas, cabeçalhos, imagens, parágrafos e outras características que não comprometam a experiência da leitura; (2) a segunda tem o objetivo de padronizar e manter apenas o texto que irá para a etapa de classificação textual, não abordada neste artigo. A divisão dessa atividade é importante para que seja mantida a originalidade da matéria, permitindo uma leitura completa e concisa pelo auditor e, quando necessário, a padronização do texto para classificação textual, como será detalhado a seguir.

Na primeira etapa, são removidos da matéria *tags* HTML que não são necessárias, *tags* que não possuem conteúdos, classes de estilo (CSS), códigos, scripts, javascripts, dentre outros. Essa etapa garante a harmonia para a apresentação da matéria para os auditores, mantendo os padrões gráficos, tais como de cabeçalhos, subtítulos e imagens. O resultado dessa etapa é apresentado na Figura 1, a seguir.

Figura 1. Comparativo entre a matéria original e a disponível no Sussurro

3.3. Classificação dos textos

Inicialmente, a classificação está baseada em matérias de treinamento com indícios de auditorias, compondo um classificador Bayesiano.

4. Resultados

Considerando o período de 01/01/2013 até 01/02/2023 (10 anos e um mês), o total de matérias coletadas neste range temporal é de 953.440, sendo 56.053 referentes a casos que podem direcionar uma auditoria na área de saúde.

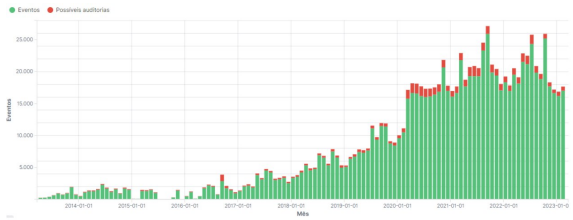

Figura 2. Série histórica gerada pelo Sussurro, com a quantidade de eventos e possíveis auditorias

A Figura 2 destaca a proporção entre as matérias publicadas e as que correspondem à possibilidade de auditoria na área de saúde, demonstrando que em todos os meses há a presença de eventos sobre o contexto deste trabalho.

5. Considerações finais e trabalhos futuros

O trabalho foi concebido a partir da constatação de que o atual quantitativo de auditores na AudSUS (Ministério da Saúde) requer esforços sobre-humanos para encerramento da fila de processos relacionados ao sistema de saúde público brasileiro. Além disto, estes auditores buscam, em diversas fontes disponíveis na internet, matérias que possibilitem a construção das trilhas de auditoria e do conhecimento necessário para execução das atividades presentes na fase analítica do processo. Essas consultas, via de regra, são realizadas por meio de métodos manuais.

Para auxiliar o processo de consulta às fontes de matérias textuais, este trabalho apresentou uma prova de conceito de uma plataforma capaz de coletar e disponibilizar o conteúdo necessário para que os auditores possam, de forma mais direta, buscar o que precisam em um local centralizado. O vídeo demonstrativo está disponível em `https://bit.ly/sussuvideo`, bem como sua utilização pode ser feita a partir do endereço `https://bit.ly/sussulais`.

Como trabalhos futuros, estima-se a evolução dos robôs para um aumento na precisão da extração do trecho que contém o texto das matérias e para a inserção de novas fontes de jornais online e redes sociais. Com relação à aplicação de NLP, ocorrerá a comparação de algoritmos de classificação de texto mapeados em revisão sistemática da literatura, a qual elencará os mais utilizados, eficientes e eficazes, no contexto de notícias e de relatórios de auditoria, além de elencar os algoritmos mais efetivos no reconhecimento de entidades nomeadas. Por fim, serão utilizados algoritmos de construção de sumários sobre os assuntos da saúde, automatizando a descoberta de novos assuntos em auditoria, baseada no aprendizado dos dos eventos que são inseridos na plataforma.

Referências

Aquino, M. (2022). Com atual efetivo, auditorias no farmácia popular demorariam 20 anos.

Brasil (1988). Constituição da república federativa do brasil.

Brasil (1993). Lei nº 8.689, de 27 de julho de 1993.

Brasil (2010). Política nacional de gestão de tecnologias em saúde.

Brasil (2015). Portaria nº 589.

Brasil (2017). Princípios, diretrizes e regras da auditoria do sus no âmbito do ministério da saúde [recurso eletrônico.

Brasil (2023). Decreto 11.358 de 1º janeiro de 2023.

Costa, T. D., e. a. (2021). Análise do perfil das ações de auditoria realizadas a partir do sistema de auditoria do sistema Único de saúde. *Revista de Administração em Saúde*.

Diakopoulos, N. (2020). Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism. *Digital Journalism*.

Lima, H. S. C., e. a. (2019). Sus, saúde e democracia: desafios para o brasil. *Ciência e saúde coletiva*.

Weichselbraun, A., e. a. (2020). Classifying news media coverage for corruption risks management with deep learning and web intelligence. *WIMS20*.