doi: 10.1590/0100-3984.2023.0096

O processamento de língua natural permite a classificação correta de laudos radiológicos em doenças benignas da vesícula biliar

Natural language processing in the classification of radiology reports in benign gallbladder diseases

Lislie Gabriela Santin[1,a], Henrique Min Ho Lee[1,b], Viviane Mariano da Silva[1,c], Ellison Fernando Cardoso[1,d], Murilo Gleyson Gazzola[1,2,e]

1. Hospital Israelita Albert Einstein, São Paulo, SP, Brasil. 2. Universidade Presbiteriana Mackenzie, São Paulo, SP, Brasil.

Correspondência: Lislie Gabriela Santin. Rua Padre Lebret, 801, Jardim Leonor. São Paulo, SP, Brasil, 05653-160. E-mail: lisliesantin@hotmail.com.

a. https://orcid.org/0000-0001-7418-746X; b. https://orcid.org/0000-0002-1266-0095; c. https://orcid.org/0009-0007-8688-7122; d. https://orcid.org/0000-0002-5542-4527; e. https://orcid.org/0000-0002-0773-6251.

Submetido em 30/8/2023. Revisado em 8/1/2024. Aceito em 19/2/2024.

Como citar este artigo:
Santin LG, Lee HMH, Silva VM, Cardoso EF, Gazzola MG. O processamento de língua natural permite a classificação correta de laudos radiológicos em doenças benignas da vesícula biliar. Radiol Bras. 2024;57:e20230096.

Resumo **Objetivo:** Desenvolver uma aplicação de processamento de linguagem natural capaz de identificar automaticamente doenças cirúrgicas benignas da vesícula biliar a partir de laudos radiológicos.

Materiais e Métodos: Desenvolvemos um classificador de texto para classificar laudos como contendo ou não doenças cirúrgicas benignas da vesícula biliar. Selecionamos aleatoriamente 1.200 laudos com descrição da vesícula biliar de nosso banco de dados, incluindo diferentes modalidades. Quatro radiologistas classificaram os laudos como doença benigna cirúrgica ou não. Duas arquiteturas de aprendizagem profunda foram treinadas para a classificação: a rede neural convolucional (*convolutional neural network* – CNN) e a memória longa de curto prazo bidirecional (*bidirectional long short-term memory* – BiLSTM). Para representar palavras de forma vetorial, os modelos incluíram uma representação Word2Vec, com dimensões variando de 300 a 1000. Os modelos foram treinados e avaliados por meio da divisão do conjunto de dados entre treinamento, validação e teste (80/10/10).

Resultados: CNN e BiLSTM tiveram bom desempenho em ambos os espaços dimensionais. Relatamos para 300 e 1000 dimensões, respectivamente, as pontuações F1 de 0,95945 e 0,95302 para o modelo CNN e de 0,96732 e 0,96732 para a BiLSTM.

Conclusão: Nossos modelos alcançaram alto desempenho, independentemente de diferentes arquiteturas e espaços dimensionais.

Unitermos: Processamento de linguagem natural; Redes neurais de computação; Aprendizado profundo; Máquina de vetores de suporte; Inteligência artificial.

Abstract **Objective:** To develop a natural language processing application capable of automatically identifying benign gallbladder diseases that require surgery, from radiology reports.

Materials and Methods: We developed a text classifier to classify reports as describing benign diseases of the gallbladder that do or do not require surgery. We randomly selected 1,200 reports describing the gallbladder from our database, including different modalities. Four radiologists classified the reports as describing benign disease that should or should not be treated surgically. Two deep learning architectures were trained for classification: a convolutional neural network (CNN) and a bidirectional long short-term memory (BiLSTM) network. In order to represent words in vector form, the models included a Word2Vec representation, with dimensions of 300 or 1,000. The models were trained and evaluated by dividing the dataset into training, validation, and subsets (80/10/10).

Results: The CNN and BiLSTM performed well in both dimensional spaces. For the 300- and 1,000-dimensional spaces, respectively, the F1-scores were 0.95945 and 0.95302 for the CNN model and 0.96732 and 0.96732 for the BiLSTM model.

Conclusion: Our models achieved high performance, regardless of the architecture and dimensional space employed.

Keywords: Natural language processing; Neural networks, computer; Deep learning; Support vector machine; Artificial intelligence.

INTRODUÇÃO

As doenças benignas da vesícula biliar são altamente prevalentes na população, podendo variar entre diversos fatores etiológicos como: colelitíase, microlitíase, adenomiomatose, polipose, colesterolose, entre outros. Desses fatores, a colelitíase é seu representante mais notório e pode ser definido como a presença de cálculos no interior da vesícula biliar, chegando a acometer cerca de 6,3 milhões de homens e 14,2 milhões de mulheres entre 20 e 74 anos nos Estados Unidos[1]. A maioria dos indivíduos é assintomática e seu diagnóstico é realizado por meio de exames de imagem de rotina ou durante a investigação de outras doenças abdominais, sendo muitos desses pacientes assintomáticos e sem necessidade de tratamento. Porém, nos

casos indicados, é realizada colecistectomia, que, apesar de definitiva, não é isenta de riscos[2]. Por esse motivo, a indicação cirúrgica necessita apresentar critérios precisos. Nesse contexto, é necessária a avaliação dos laudos radiológicos, a fim de obter as informações sobre as doenças da vesícula biliar. Essa avaliação, realizada de forma manual individual, é altamente trabalhosa quando considerada em grande escala. Para solucionar esse problema, foi utilizado o processamento da língua natural (PLN).

O PLN é um campo de pesquisa aplicada interdisciplinar que inclui ciência da computação e inteligência artificial que analisa dados de linguagem natural, servindo como uma interseção entre ciência da computação e linguística, com o intuito de desenvolver um sistema de suporte à decisão. Os sistemas para o suporte à decisão clínica são quaisquer *softwares* projetados para auxiliar diretamente na tomada de decisão clínica, em que as características de cada paciente são combinadas a uma base de conhecimento, com o objetivo de gerar avaliações ou recomendações específicas que são apresentadas aos profissionais de saúde para sua consideração, podendo incluir nos modelos avançados, por exemplo, verificação de interações medicamentosas e doenças, suporte de dosagem individualizada durante insuficiência renal, recomendações sobre testes laboratoriais durante o uso de drogas, entre outros[3].

Os métodos do PLN fornecem os meios para que as pessoas possam trabalhar usando sua língua natural e ainda ser possível desenvolver algoritmos que manipulam, aumentam e transformam a língua natural em um formato computável. Dessa forma, se provou eficaz na extração de informações de relatórios de radiologia, incluindo detecção de achados críticos, avaliação de qualidade e geração de anotações e conjuntos de dados[4], sendo, portanto, o método ideal para desenvolvimento do nosso algoritmo.

MATERIAIS E MÉTODOS

O desenvolvimento do presente trabalho utilizou como base para o projeto 1.100 laudos de exames de imagem que descreviam alterações da vesícula biliar realizados na instituição, durante os meses de janeiro e fevereiro de 2018. Também foram incluídos nessa base, 100 laudos de exames de pacientes escolhidos aleatoriamente no mesmo período e que realizaram acompanhamento clínico por colecistectomia, totalizando 1.200 laudos ao todo. O escopo de laudos utilizados é composto por três tipos diferentes de exames, sendo a distribuição apresentada na Tabela 1. O conjunto de laudos disponibilizados foi dividido em dois grupos, possuindo cada um 600 laudos, visando à etapa posterior de anotações realizadas por radiologistas.

Pré-processamento

A correta identificação dos termos mediante utilização do modelo com *named entity recognition* necessita que o texto esteja padronizado em sentenças completas e únicas. Com base nessa necessidade, foi observado que parte dos laudos apresentava desconfiguração de frases em que era possível identificar quebra de linha em meio à sentença, múltiplas sentenças na mesma linha e casos especiais de pontuação para divisão de sentenças.

Tabela 1—Tipos de laudos utilizados e sua distribuição nos grupos de anotação.

Tipos de laudos	Grupo 1	Grupo 2
Ultrassonografia	318	300
Tomografia computadorizada	201	204
Ressonância magnética	81	96
Total	600	600

Os textos completos dos laudos disponibilizados foram fracionados em linhas, utilizadas para treinamento de modelo de sentenciador. As linhas foram rotuladas em quatro classes: 1) sentenças únicas corretas; 2) sentença única com quebra indevida; 3) sentenças múltiplas completas; 4) sentenças múltiplas com quebra.

A etapa de pré-processamento dos dados, indicada como *data cleaning*, é composta por um total de quatro passos principais. O primeiro tipo de tratamento nos dados consiste na conversão em letras minúsculas, seguido por remoção de pontuação e *stop words*. Posteriormente, o *corpus* passa por duas conversões para representação quantitativa e numérica, que é o *bag of words* e o *term frequency-inverse document frequency*. Além disso, foi realizada a remoção de dados sensíveis, informações referentes a paginação, instruções de acesso ao laudo e identificação do profissional responsável pela revisão e liberação do documento. Essas sentenças presentes em todos os laudos não influenciam de nenhuma forma nos achados clínicos, portanto, a remoção dessas partes do texto somente colabora na redução da dimensão do dado utilizado.

Anotações de laudos

A construção dos laudos radiológicos é composta por diversas constatações referentes às condições fisiológicas e estruturais dos órgãos, porém, nem todas essas informações contribuem diretamente para o diagnóstico final. A necessidade de uma intervenção médica pode ser observada a partir da presença de algumas expressões ou sentenças nos laudos (Tabela 2) relacionados aos rótulos selecionados. Todos esses rótulos selecionados foram codificados segundo o RadLex, ontologia específica utilizada na radiologia[5].

Tabela 2—Termos chaves anotados nos laudos.

RadLex code	Termo
RID187	Gallbladder
RID3394	Cholecystitis
RID34607	Microlithiasis
RID3869	Adenomyomatosis
RID3881	Polyp
RID4989	Gallstone
RID5198	Porcelain gallbladder
RID5215	Cholesterolosis

Todos os documentos foram anonimizados e alocados em um ambiente interno dentro da rede da instituição, seguro e específico para a anotação. Três radiologistas e um residente de radiologia realizaram a análise e identificação dos termos estabelecidos, indicando a posição e a sentença da ocorrência. A ferramenta do INCEpTION (Figura 1) foi utilizada para a realização das anotações individuais de cada especialista e cada anotador possuía um *login* e senha específicos para acesso, com acesso apenas dentro do hospital. A utilização do INCEpTION permitiu também que posteriormente fosse realizada uma comparação entre os diferentes achados em cada documento para uma definição final dos termos do laudo.

Tendo em vista a pouca experiência dos médicos radiologistas com esse trabalho de anotação, foi elaborado um manual com regras claras sobre como entrar no ambiente INCEpTION, que palavras incluir na anotação e como classificar corretamente os termos (Figura 2), visando a responder possíveis dúvidas e uniformizar os dados para melhor treinamento da máquina.

Além das anotações dos termos clínicos, os laudos foram classificados de acordo com a presença de indicações cirúrgicas de colecistectomia. Para tanto, foi elaborada a Tabela 3 por um conselho de especialistas da instituição, com base em *guidelines* internacionais e do próprio hospital[6,7], que apresentavam indicações de colecistectomia na sua prática clínica e cirúrgica, sendo usada para rotulação como em conformidade ou não para indicação cirúrgica.

Tabela 3—Tabela de classificação de conformidade e não conformidade na indicação de colecistectomia, elaborada por especialistas.

Aspectos de conformidade cirúrgica	Aspectos de possível não conformidade cirúrgica
Cálculo biliar	Adenomiose sem sintomas ou critérios associados
Microlitíase	Colesterolose sem sintomas ou critérios associados
Barro biliar/bile espessa	Laudo sem alteração na vesícula biliar
Colelitíase/colecistite	Ausência de exames
Pólipo ≥ 5 mm	—
Vesícula em porcelana	—

Curadoria

Após a anotação dos laudos, foi realizada uma fase de curadoria a fim de estabelecer um *corpus* padrão ouro. Nesta fase, as anotações foram comparadas entre as duplas, resolvendo os pontos de divergência levantados por meio de consenso entre os participantes do trabalho. A concordância entre os anotadores foi avaliada por meio de um coeficiente de concordância de kappa de Cohen, que permite avaliar a complexidade da tarefa e a precisão como a anotação foi feita. Um valor de concordância acima de 0,81 é considerado perfeito[8]. Foram obtidos no grupo 1 e grupo 2 valores respectivos de 0,97 e 0,88. O padrão ouro utilizado para o treinamento e teste dos modelos foram as anotações resultantes da fase de curadoria.

Treinamento

Para o treinamento utilizamos técnicas como o *Word embeddings*, que transformam palavras em vetores de números reais, representando-as em um espaço n-dimensional, normalmente 300, e aprendendo com grandes *corpora* não anotados, sendo capazes de capturar conhecimento sintático, semântico e morfológico, possibilitando

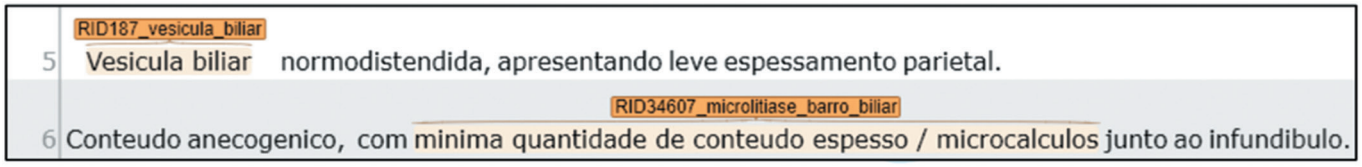

Figura 1. Ambiente de anotação no INCEpTION com algumas entidades já classificadas pelo radiologista.

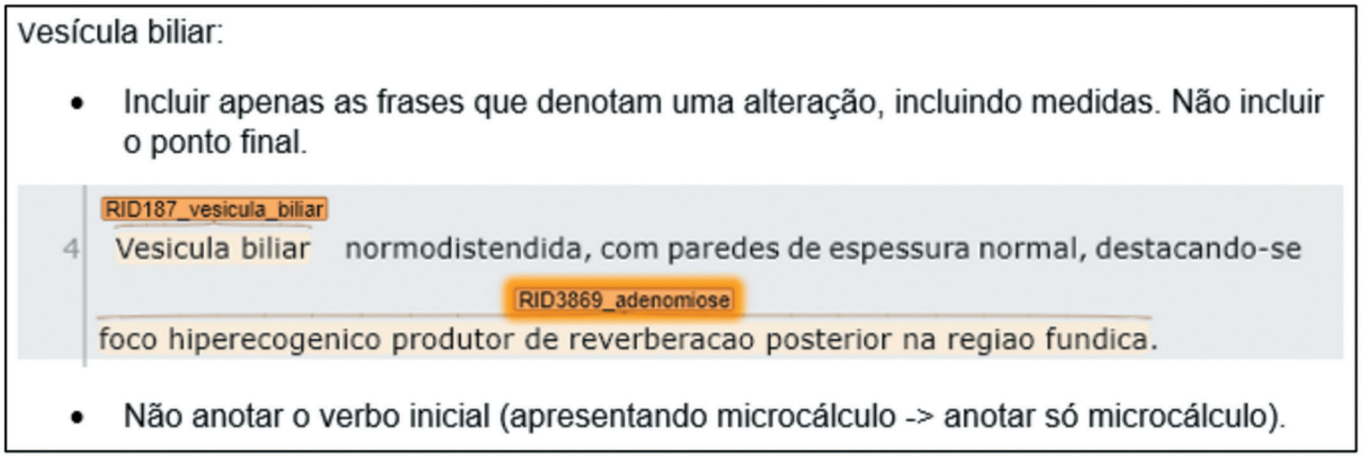

Figura 2. Exemplos de regras para resolução de possíveis dúvidas descritas no manual para uniformização das anotações.

eficientemente a "tradução" dos textos e tornando-se comum em sistemas de PLN.

Dentro do *Word embeddings* existem diversos modelos disponíveis, dentre eles o Word2Vec, que foi um dos primeiros a ganhar popularidade[9]. Este modelo usa uma rede neural de duas camadas que processa texto vetorizando as palavras. Sua entrada é um *corpus* de texto e sua saída é um conjunto de vetores, um de cada palavra. Esses vetores agrupam as palavras semelhantes, permitindo a verificação da semelhança matemática dos vetores, fazendo uso da similaridade do cosseno[9]. Com dados e textos corretos e suficientes, o Word2Vec consegue fazer suposições altamente precisas sobre o significado de uma palavra com base na sua ocorrência no *corpus*, por isso foi o método utilizado para treinamento do nosso modelo.

Algoritmos de *deep learning* foram adotados, sendo a arquitetura de rede neural convolucional (*convolutional neural network* – CNN), adaptada para texto, a primeira utilizada. A construção dessa rede contou com duas camadas convolucionais contendo 64 e 128 filtros com janelas convolucionais de dimensão 7 e 5, respectivamente. A função de ativação *tanh* foi utilizada em ambas as camadas convolucionais, seguida por *max pooling* e outras duas camadas densas com ativação Relu intermediadas por uma camada de *dropout*. O otimizador selecionado para utilização foi Adam, com uma *learning rate* de 0.0001.

As arquiteturas de redes neurais recorrentes também são amplamente utilizadas no contexto de PLN, enfatizando a *long short-term memory* (LSTM). A variação bidirecional desse tipo de arquitetura (BiLSTM) foi utilizada para treinamento, de modo que duas camadas BiLSTM, contendo 256 e 128 neurônios respectivamente, foram intermediadas por uma camada de *attention* com ativação *sigmoid*. A inclusão de uma *dense layer* também foi realizada após a última camada BiLSTM, com ativação Relu. Em ambos os tipos de arquitetura, foi realizada a inclusão de uma camada de *Word embedding* para a vetorização do texto. Na língua inglesa existem diversos repositórios de processamento de língua natural voltado para a área da saúde, sendo o bioNER[10] um exemplo. Na língua portuguesa, porém, houve dificuldades, por não haver esse banco de dados específico, sendo utilizado nos experimentos o NILC-Embedding[11] em português do Brasil com 300 dimensões e 1000 dimensões.

O treinamento foi realizado usando uma máquina com processador Intel Core i5-10210U CPU @ 1.60GHz 2.10 GH e 16GB RAM, e a linguagem utilizada foi o Python 3.8.

RESULTADOS

A validação do desempenho dos modelos de CNN e BiLSTM foram executados separando o *dataset* em conjuntos de treinamento, validação e teste na proporção 80%, 10% e 10%.

Os modelos de CNN com lematização apresentaram F-score de 0,95945 e 0,95302, considerando Word2Vec de 300 e 1000 dimensões, respectivamente. Já as arquiteturas de CNNs recorrentes do tipo BiLSTM atingiram o F-score de 0,96732 para Word2Vec 300 dimensões e de 0,96732 para 1000 dimensões (Tabela 4).

Tabela 4—Resultados da CNN e BiLSTM F-score (80/10/10).

Modelo	F-score (80/10/10)
CNN + lematização + Word2Vec 300d	0,95945
CNN + lematização + Word2Vec 1000d	0,95302
BiLSTM + lematização + Word2Vec 300d	0,96732
BiLSTM + lematização + Word2Vec 1000d	0,96732

80/10/10, distribuição proporcional dos subconjuntos de treinamento, validação e teste, respectivamente; 300d, espaço vetorial de 300 dimensões; 1.000d, espaço vetorial de 1.000 dimensões.

DISCUSSÃO

Dentre as diferentes arquiteturas de *deep learning*, foram treinadas para a tarefa de classificação as CNNs e as recorrentes. Em cada um dos casos, uma camada de *embedding* Word2Vec, pré-treinada com *corpus* em português, foi alocada à entrada de modo a representar o texto de forma vetorial. A *Word embedding* foi variada entre 300 e 1000 dimensões, visando a estabelecer também sua influência nos resultados.

As métricas obtidas por meio da validação dos diferentes modelos mostram desempenho semelhante. A ampliação de dimensões na representação Word2Vec parece não ter impacto significativo, visto que os resultados foram semelhantes. Uma possível causa para esse comportamento pode ser o escopo fechado da avaliação de uma estrutura anatômica específica, e de doenças benignas cirúrgicas, de modo que o número de palavras contidas em uma frase e suas variações seja bem limitado. Os laudos radiológicos possuem descritores técnicos específicos para facilitar a comunicação entre os profissionais da área de saúde, de modo que isso encurte o tamanho dos textos médicos.

Embora tenhamos obtido resultados positivos, é crucial identificar e analisar as limitações que podem impactar a interpretação e aplicação desses resultados. Uma limitação reside na exclusividade da origem dos laudos, provenientes de um único hospital. Isso limita a representatividade, pois diferentes instituições de saúde podem variar em termos de linguagem e abordagens clínicas. Além disso, os laudos não foram originalmente destinados ao treinamento do algoritmo, mas sim ao diagnóstico de doenças na rotina clínica. Isso pode afetar a qualidade dos dados, pois podem não incluir a diversidade necessária para um treinamento abrangente, proporcionando uma visão limitada da aplicabilidade do algoritmo, especialmente em situações clínicas menos comuns ou especializadas. A decisão de focar apenas em doenças benignas da vesícula biliar foi tomada em razão de restrições quanto à complexidade associada à avaliação de prontuários e confirmação das doenças diagnosticadas. No entanto, essa escolha pode limitar a aplicabilidade do algoritmo em casos de doenças

mais raras e complexas, comprometendo a generalização dos resultados. Ao reconhecer e discutir essas limitações de maneira transparente, contribuímos para a integridade do estudo e promovemos o avanço contínuo da aplicação de algoritmos em contextos clínicos.

CONCLUSÃO

Doenças da vesícula biliar são bastante prevalentes na população e seu tratamento comumente envolve realização de procedimento cirúrgico, que apesar de raramente apresentar complicações, não é isento de riscos. Para garantir que a indicação cirúrgica seguisse critérios precisos, desenvolvemos um protótipo de um o sistema de suporte à decisão clínica, que extrai informações de laudos radiológicos, classificando doenças da vesícula biliar em cirúrgicas ou não cirúrgicas. Para facilitar esse processo, foi utilizado o PLN. Este permite a automação de uma diversidade de tarefas em radiologia e é uma área de pesquisa valiosa na análise, agregação e simplificação de dados não estruturados (textual), já tendo demonstrado potencial significativo na análise de laudos radiológicos.

Existem inúmeras bibliotecas e ferramentas de código aberto disponíveis que facilitam sua aplicação em benefício da radiologia. Radiologistas que entendem suas limitações e potencial estarão mais bem posicionados para avaliar modelos de PLN, entender como eles podem melhorar o fluxo de trabalho clínico e facilitar esforços de pesquisa envolvendo grandes quantidades de linguagem humana. Por sua vez, a radiologia tem um potencial significativo para se beneficiar da capacidade da PLN de converter laudos de radiologia em dados legíveis por máquina.

Nossos modelos obtiveram alto desempenho para identificar e extrair automaticamente a presença ou ausência de doenças cirúrgicas benignas da vesícula biliar a partir de laudos radiológicos usando o PLN, independentemente de diferentes arquiteturas e espaços dimensionais. Essas abordagens podem ser estendidas a outros cenários clínicos usando método semelhante para extrair e estruturar informações de grandes conjuntos de dados.

REFERÊNCIAS

1. Afdhal NH, Zakko SF. Gallstones: epidemiology, risk factors and prevention. UpToDate [Internet]. [cited 2022 Sep 27]. Available from: https://www.uptodate.com/contents/gallstones-epidemiology-risk-factors-and-prevention?search=Gallstones%3A%20epidemiology%2C%20risk%20factors%20and%20prevention&source=search_result&selectedTitle=1~150&usage_type=d.
2. Zakko SF. Overview of nonsurgical management of gallbladder stones. UpToDate [Internet]. [cited 2022 Oct 3]. Available from: https://www.uptodate.com/contents/overview-of-nonsurgical-management-of-gallbladder-stones?search=Overview%20of%20nonsurgical%20management%20of%20gallbladder%20stones&source=search_result&selectedTitle=1%7E150&usage_type=default&display_rank=1.
3. Wasylewicz ATM, Scheepers-Hoeks AMJW. Clinical decision support systems. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science [Internet]. Springer Cham; 2019. [cited 2023 Aug 30]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK543516/.
4. Bressem KK, Adams LC, Gaudin RA, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics. 2021;36:5255–61.
5. RSNA. RadLex radiology lexicon [Internet]. [cited 2022 Sep 27]. Available from: https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon.
6. Sociedade Beneficente Israelita Brasileira. Colecistectomia laparoscópica [Internet]. [cited 2022 Oct 3]. Available from: https://medicalsuite.einstein.br/pratica-medica/Pathways/Pathway%20Colecistectomia.pdf.
7. Soper NJ. Laparoscopic cholecystectomy. UpToDate [Internet]. [cited 2023 Nov 26]. Available from: https://www.uptodate.com/contents/laparoscopic-cholecystectomy?search=Laparoscopic%20cholecystectomy&source=search_result&selectedTitle=1%7E75&usage_type=default&display_rank=1.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
9. Hartmann NS. Adaptação lexical automática em textos informativos para o ensino fundamental [tese]. São Paulo, SP: Universidade de São Paulo; 2020. [cited 2022 Sep 27]. Available from: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-29072020-161751/.
10. Wang K, Zhang Y, Ren S, et al. Cross-type biomedical named entity recognition with deep multi-task learning [Internet]. [cited 2022 Oct 10]. Available from: https://xuanwang91.github.io/BioNER/.
11. Repositório de Word Embeddings do NILC [Internet]. [cited 2022 Oct 4]. Available from: http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc.