Avaliação de modelos para extração de dados não estruturados de um sistema EHR para atender a estrutura final de uma ontologia

Diego Pinheiro da Silva[1], **Blanda Helena de Mello**[1], **William da Rosa Fröhlich**[1], **Sandro José Rigo**[1], **Marco Antonio Schwertner**[1], **Cristiano André da Costa**[1]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos, 950 – Cristo Rei, São Leopoldo – RS, 93022-750 – Brazil

Abstract. *There is a significant increase in the number of Electronic Health Records (EHRs) that accommodate unstructured data in natural language. Manual analysis of this data is not feasible due to the large volume, whose tendency is to continue to increase. In this scenario, there is a need for approaches that allow the structuring of this information to help health professionals in data analysis, treatment indication, disease diagnosis, and others. This research aims to develop a model for processing unstructured data from EHRs, observing the objective of representing them in ontology structures. Preliminary experiments were carried out and indicated promising results for the model.*

Resumo. *Há um aumento significativo no número de Electronic Health Records (EHRs) que acomodam dados não estruturados em linguagem natural. A análise manual destes dados é inviável devido ao grande volume existente, cuja tendência é continuar a aumentar. Há uma necessidade de abordagens que permitam a estruturação destas informações para que possam auxiliar os profissionais de saúde na análise dos dados, indicação de tratamento, diagnóstico de doenças, entre outros. Esta pesquisa tem como objetivo desenvolver um modelo para processamento de dados não estruturados de EHRs observando o objetivo de representá-los em estruturas de ontologias. Experimentos preliminares foram realizados e indicaram resultados promissores para o modelo.*

1. Introdução

A crescente adoção de prontuários em formato digital que acomodam dados não estruturados, com textos e observações em Linguagem Natural (LN) [Ngiam and Khor 2019], aliada a adoção de *Electronic Health Records* (EHR) e o consequente aumento no volume de dados não estruturados disponíveis, têm sido motivos de interesse para pesquisas que buscam o tratamento da LN [Dhole and Uke 2014]. As informações médicas historicamente foram extraídas dos registros do paciente por especialistas clínicos. Essa abordagem tem limitações claras de escalabilidade e tempo, além de custosa [Koleck et al. 2019]. A disponibilidade de EHRs para reutilização de dados secundários criou uma oportunidade para exploração de Processamento de Linguagem Natural (PLN) em narrativas de texto livre, buscando soluções de apoiar aos profissionais de saúde [Kreimeyer et al. 2017]. Os dados contidos no EHR, como relatórios clínicos e observações médicas, podem ser usados para registros de doenças, estudos epidemiológicos, vigilância de segurança de medicamentos, ensaios clínicos, auditorias de saúde e muito mais [Ford et al. 2016]. Entretanto, torna-se inviável a análise manual desses dados devido ao grande volume de dados não estruturados existente e gerados diariamente [Ngiam and Khor 2019].

Neste cenário, a Inteligência Artificial (IA) vem sendo utilizada de forma crescente na área médica. Em particular, abordagens que combinam *Machine Learning* (ML) à prática clínica, com aplicações para processamento de dados pré-clínicos, assistência a diagnósticos, tomada de decisão para tratamento e alerta precoce como parte da prevenção primária e secundária [Adlung et al. 2021]. Esta abordagem permite o reconhecimento de padrões e, com a abundância de conjuntos de dados disponível, a aplicação para extrair informações relevantes dos dados como no domínio da saúde [Mahesh 2020]. De forma mais recente, observa-se a escalada do uso de abordagens de *Deep Learning* (DL), possibilitando resultados promissores para tratamento de diferentes tarefas, desde a análise de imagens até extração de informações de texto [Cuocolo et al. 2020]. Por outro lado, observam-se trabalhos integrando recursos como extração de informação, uso de ontologias de saúde, aplicação de modelos vetoriais de linguagem em textos médicos, uso de Redes Neurais Artificiais (RNA) e sistemas de recomendação em saúde [Lee et al. 2016], [Yang et al. 2019] e [Christopoulou et al. 2020].

Este artigo tem como objetivo descrever o modelo conceitual em desenvolvimento para extração de informação em dados não estruturados em saúde para população de uma ontologia, e avaliar modelos para reconhecimento e extração de informações. Os resultados podem ser aplicados no acompanhamento automatizado e para subsídios ao diagnóstico clínico, análise de qualidade, projeção de cenários para orçamento, identificação de epidemias, predição, recomendação de diagnósticos ou protocolos. Portanto, há contribuições no nível da aplicação em sistemas EHR para oncologia, ampliando a capacidade desses sistemas para o uso dos dados não estruturados, e também com a experimentação e avanços para o reconhecimento de entidades e extração de relações, com futura avaliação de um modelo BERT para verificar sua extração de informações.

2. Trabalhos Correlatos

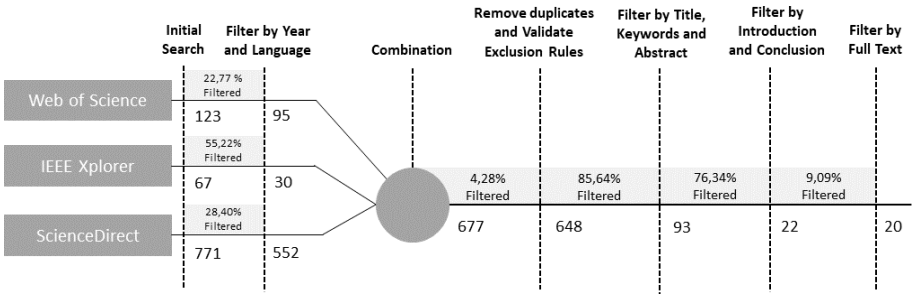
Para o estudo de trabalhos correlatos, foi realizada uma revisão sistemática de literatura na qual, após um processo de filtragem de 961 publicações, foram analisados 20 trabalhos com foco em extração de relações. O protocolo foi desenvolvido utilizando uma mescla do protocolo de [Kitchenham and Charters 2007], referente à área da computação, e o protocolo de [Prisma 2021], direcionado para à área da saúde[1]. A Figura 1 ilustra um fluxograma geral dos resultados das classificações dos artigos em cada fase de seleção.

No contexto de extração de informações, algumas abordagens envolvem a extração em sistemas EHR e classificação automática de doenças, com um percentual de 67% destas obras utilizando abordagens baseadas em regras [Ford et al. 2016]. Por outro lado, sistemas baseados em regras são limitados [Dhole and Uke 2014], sendo um dos motivos para o crescente envolvimento de PLN para identificação de padrões [Gonzalez-Hernandez et al. 2017]. Neste sentido, os modelos pré-treinados *bidirectional encoder representations from transformers* (BERT) tem apresentado resultados promissores em tarefas de PLN, para reconhecimento de entidades nomeadas ou extração de relações. Assim como a exploração destes para domínios específicos. Modelos BERT permitem a realização de *fine-tuning* em *datasets* específicos, prática que pode ser vantajosa para incrementar a eficiência em cenários de domínio específico [Li et al. 2020],

[1]Repositório GitHub `https://github.com/blandamellus/SLR_Protocol_DeepLearning2021`

[Morgan et al. 2021]. Os autores [Xue et al. 2019] também exploraram a utilização de *Transformers*, mas com foco na mineração de dados em domínio biomédico. Enquanto o modelo BERT é considerado um modelo genérico para tarefas de PLN, o modelo Bio-BERT representa o atual estado da arte para aplicação em domínio biomédico.

Figura 1. Fluxograma dos resultados das fases de seleção



O estado da arte mostra trabalhos relevantes e com bons resultados quanto ao uso do BERT. Sendo assim, serão utilizados modelos BERT e Bi-LSTM (Rede de memória de longo prazo bidirecional) para reconhecimento de entidades e extração de relações. Além disso, será realizado um *fine-tuning* utilizando notas médicas reais em português com base no *dataset* desenvolvido nessa pesquisa. Escolheu-se o BERT porque ele é amplamente adotado e constitui o estado da arte atual. Os modelos conhecidos exploram pouco os aspectos linguísticos e semânticos [Morgan et al. 2021], [Li et al. 2020] e [Schneider et al. 2020]. Embora o modelo BERT seja largamente adotado, a exploração destes em cenários da língua portuguesa ainda apresentam uma carência de ferramentas e mesmo limitações quanto a datasets disponíveis. Esse cenário por vezes é solucionado aplicando a tradução dos dados como parte do pipeline, para usufruir do ferramental disponível para língua inglesa. Esta prática pode acarretar a perda de contexto ou significados e mesmo eficiência. Neste sentido, [Schneider et al. 2020] propõe transferir as informações aprendidas do modelo BERT-multilíngue para um *corpus* de artigos científicos biomédicos e narrativas clínicas em Português Brasileiro. O modelo Bio-BERTpt foi avaliado com experimentos NER, em dois *corpus* anotados contendo narrativas clínicas e os resultados foram comparamos com modelos BERT existentes.

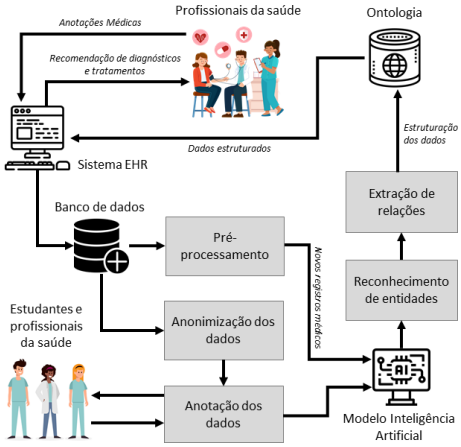
Dessa forma, destaca-se como lacunas os seguintes pontos. O primeiro trata-se da língua portuguesa, que possui poucos recursos atualmente. O segundo, os tipos de dados a serem trabalhados, sendo que a maior parte dos trabalhos usam dados fora da área da saúde e que não representam/originam de contextos reais. O terceiro ponto está relacionado com as técnicas utilizadas, sendo que ainda poucos trabalhos exploram o uso do BERT e *Transformers*, que vem ganhando destaque no estado da arte atual, para tarefas de extração. O quarto ponto trata da estruturação de dados com base na percepção da restrita utilização de recursos para estruturação de dados, tais como ontologias.

3. Materiais e métodos

Neste artigo se objetiva experimentar técnicas de DL, em especial o modelo BERT, *Transformers* e Bi-LSTM, para reconhecimento de entidades e extração de relações em dados oncológicos gerados em um sistema EHR. Observando este cenário, a revisão sistemática evidenciou uma lacuna de pesquisa quanto ao uso de dados em português, principalmente

se tratando de *Transformers* e BERT para extração de informações, visto que são técnicas recentes e foram pouco mencionadas na literatura. Para a realização do estudo, conta-se com a parceria da empresa, a qual possui um sistema EHR de oncologia denominado Sistema Gemed Oncologia (SGO). A empresa trabalha com um grupo de clinicas e hospitais, fornecendo a solução para registro eletrônico dos dados de atendimentos médicos, de enfermagem, farmacêuticos, entre outras especialidades. Portanto, esta pesquisa aplicará um estudo de caso na área da oncologia, utilizando dados e outros recursos do sistema. A Figura 2 apresenta essa visão contendo os elementos presentes dentro desse contexto, que considera clínicas médicas e expressa as necessidades dos profissionais de saúde.

Figura 2. Visão geral do contexto da abordagem pesquisa



O processo considera a criação de um registro médico ou anotação de um profissional de saúde em um sistema EHR. Nos casos observados, essa situação gera registros médicos em formato textual não estruturado e informações clínicas estruturadas. **Profissionais da saúde e anotações médicas:** a aplicação da pesquisa possibilitará atender e assistir profissionais da saúde, sendo eles médicos, enfermeiros e farmacêuticos, que atendem pacientes no campo da oncologia em clínicas e hospitais de diversos pontos do Brasil. A cada consulta, um especialista registra uma evolução do paciente, onde são preenchidas informações relevantes para o tratamento do paciente, essas observações podem ser compostas de texto livre e dados estruturados. Após a conclusão (e em paralelo) a essa etapa, as anotações médicas são inseridas em um sistema EHR. **Sistema EHR de oncologia:** estudo de caso desta pesquisa, será utilizado o SGO como Sistema EHR. No sistema, as notas médicas contêm eventos clínicos registrados por médicos, enfermeiras e outros profissionais de saúde, com informações clínicas sobre o paciente. Para registrar essas informações, o profissional de saúde usa um formulário personalizável do sistema para cada evento clínico. O banco de dados do SGO armazena todas as informações sobre a saúde dos pacientes, como anotações médicas, prescrições eletrônicas e histórico de consultas dos pacientes. Para cada clínica de oncologia que utiliza o sistema, é criada uma instância do banco de dados para armazenar suas informações separadamente.

Em relação a **Exportação, Anonimização e Anotação dos dados**, esse processo será exclusivamente realizado para experimentos, visto que em produção os novos registros médicos passarão por um pré-processamento e serão utilizados pelo modelo de IA, conforme mostra na Figura 2. A anonimização é o processo que remove a associação entre o conjunto de dados de identificação e o titular dos dados. A pseudo-anonimização é

um tipo específico de anonimato que remove a associação com um titular dos dados e adiciona uma associação entre um determinado conjunto de características relacionadas ao titular dos dados e um ou mais pseudônimos. A pseudo-animização pode ser recuperável ou irrecuperável, ou seja, com ou sem a possibilidade de re-identificar o titular dos dados. Nessa pesquisa, será aplicado um processo de anonimização para evitar a divulgação e identificação única de pacientes e profissionais de saúde. Após fnalizado, os dados estão prontos para anotação e geração do *dataset*.

A Anotação é o processo de rotulagem de dados. Essa pesquisa contará com o apoio de especialistas na área da saúde da Unisinos, que utilizarão os dados extraídos do banco de dados e, com o software UBiAI[1], anotarão os dados clínicos conforme os rótulos previamente definidos. Após essa etapa, será possível construir o dataset para treinamento do modelo proposto. o qual denominaremos nesse trabalho de *Dataset* Dataset Gemed Onco (DGO).Este *dataset* será criado a partir do acesso aos dados da área de saúde em oncologia. Os *datasets* utilizados serão disponibilizados pela empresa Interprocess, em parceria com este projeto. Os dados serão reais e não-estruturados. As informações serão extraídas de textos livres usadas para treinar os modelos de classificação de texto. As informações dos pacientes que serão selecionadas são o diagnóstico de CDI e o protocolo de quimioterapia utilizado no tratamento do paciente. Com o *dataset* anonimizado e anotado, os modelos de DL estarão prontos para serem treinados.

Na sequência, propõe-se a **Identificação de entidades e extração de relações**, nessa etapa, visa-se desenvolver um modelo baseado na arquitetura *Transformer* para o reconhecimento de entidades e extração de relações entre as mesmas, no domínio de textos oncológicos. A arquitetura final do modelo será composta por dois modelos: BERT e Bi-LSTM. O BERT será treinado para efetuar o reconhecimento das entidades, assim como a extração de características que representarão o texto. Em contrapartida, a Bi-LSTM será responsável por classificar a relação existente entre as entidades reconhecidas. Existem muitos modelos BERT propostos, portanto, com o objetivo de avaliar a performance do modelo BERT, diferentes modelos iniciais serão incluídos nos experimentos e o que obtiver o melhor resultado na tarefa de reconhecimento das entidades, será empregada para a validação do modelo proposto final, para a classificação das relações com a Bi-LSTM. Os modelos selecionados inicialmente são o "BERT-base-multilingual-cased" [de Souza et al. 2021], "BioBERT-PT" [de Souza et al. 2021], [Ji et al. 2020] e "BERTimbau" [Lopes et al. 2021], e [Souza et al. 2020]. Quanto à Bi-LSTM, além de parametrizações do tamanho do modelo e quantidade de camadas, também será feito um teste com BiGRUs, para avaliar qual abordagem de Redes Neurais Recorrentes pode apresentar resultados melhores. Os modelos serão aplicados nos dados reais do DGO. Primeiramente será obtido o melhor resultado possível com o BERT para o reconhecimento das entidades, e depois o treinamento da Bi-LSTM.

Após realizada a etapa de extração de informações, tem-se a necessidade de propor um formato para representação dos dados, etapa **Estruturação da Ontologia**. Na etapa final, os dados serão estruturados em uma base de conhecimento. Para essa pesquisa, optou-se pelo uso de uma ontologia que será projetada para representar as entidades, conceitos e relações específicas identificadas na extração dos dados. O uso de padrões em saúde como OpenEHR [Alemzadeh and Devarakonda 2017] e SNOMED

[1]https://ubiai.tools/login

CT® [Bucur et al. 2013] serão explorados. Assim como observando uma forma de validar esse modelo proposto, definiu-se uma **Metodologia de avaliação**, com o objetivo de avaliar o modelo em dois contextos, na área da saúde e quanto a aspectos computacionais. Com foco na área da saúde, o modelo e seus resultados serão validados com especialistas na área da oncologia, em estudos de casos aplicados ao SGO. Serão realizados experimentos com coleta de dados em questionários e acompanhamentos com os profissionais afim de validar sua percepção quanto ao uso do modelo. No contexto computacional, serão avaliadas as respostas do modelo utilizando métricas padronizadas, tais como Precisão (*Acurracy*), *Recall*, F1 Score, Macro AVG e Weighted AVG. Além dessas métricas, os resultados serão comparados com outras pesquisas semelhantes encontradas no estado da arte. Em uma segunda avaliação, será desenvolvido um próprio modelo BERT ou *Fine-tuning* com dados oncológicos, buscando principalmente melhorar o modelo atual e gerar um melhor resultado. Estes resultados serão avaliados comparando-os com os experimentos e modelos BERT anteriores.

4. Experimentos Preliminares

Os experimentos tem o objetivo de explorar possíveis resultados que possam agregar valor a esta pesquisa, empregando DL e ML. Para tanto, foram realizados dois tipos de experimentos: (1) geração de dataset e classificação; (2) reconhecimento de entidades nomeada (REN). Para **geração dataset e classificação** utilizou uma primeira versão do DGO, com notas clínicas em português do Brasil, extraídas do sistema EHR (domínio da oncologia). Aplicou-se a tarefa REN ao *dataset* de informações de medicamentos (DID) em inglês.

No experimento (1), **Geração de dataset e Classificação**, desenvolveu-se três versões de testes do DGO. As informações foram extraídas de textos livres e foram usadas para treinar os modelos de classificação de texto e, posteriormente, para sugerir o diagnóstico de novos textos usando os classificadores treinados. As seguintes informações estruturadas do paciente foram selecionadas para esse experimento: o diagnóstico de CDI do paciente e o protocolo de quimioterapia usado no tratamento do paciente. Na etapa de **Pré-processamento do *dataset***, foram executadas as seguintes tarefas: (A) Tokenização: divide o texto em *tokens* que correspondem a palavras; (B) Filtragem de *stop-words*: remoção das palavras mais comuns no idioma brasileiro, pontuação e caracteres especiais; (C) Caixa dobrável: converte todas as palavras para minúsculas. Na sequência, a **Extração de aspectos de interesse** transforma as anotações médicas para uma representação veterial, empregando o método *Bag-of-Words* (BoW)–texto em vetores de comprimento fixo, contabilizando a frequência das palavras. Nos *datasets* por evento clínico, o BoW foi gerado para cada nota médica e, nos *datasets* por paciente, para cada paciente e respectivas anotações médicas. Essa representação nas notas médicas resultou em uma representação esparsa, a sequência vetorial continha muitos zeros. Portanto, aplicou-se a técnica de Análise de Componentes Principais (PCA) para reduzir a esparsidade dos dados. PCA converte um conjunto de observações de recursos possivelmente correlacionados em um conjunto de valores de recursos linearmente não correlacionados.

As **Arquiteturas de Machine Learning e Deep Learning** selecionadas para a tarefa de classificação de texto foram: Rede Neural Multilayer Perceptron (MLP); Regressão Logística; Classificador de Árvore de Decisão; Classificador Random Forest; Classificador Extra Tress; K-Nearest Neighbors (KNN). Os conjuntos de dados foram divididos em dois grupos para treinar e testar as redes neurais, respectivamente 80% e

20%. A divisão dos dados foi feito randomicamente. Os algoritmos de aprendizagem de máquina foram implementados usando o scikit-learn, na linguagem Python. No primeiro conjunto de experimentos, foram realizados sete testes, com os seguintes detalhes da arquitetura: MLP com uma camada oculta e 500 neurônios; MLP com duas camadas ocultas 800 e 500 neurônios; um classificador de regressão logística; uma árvore de decisão máximo de vinte níveis e três amostras por folha; uma Random Forest máximo de vinte níveis e três amostras por folha; árvores extras máximo vinte níveis e três amostras por folha; um classificador KNN com um K unitário. O *dataset* por evento da clínica pequena contém 3.308 notas clínicas e 397 pacientes distintos, usado para realizar as classificações, utilizando o conjunto de dados pré-processados para os classificadores ML. Quanto à precisão média, o Macro F1 score e o Weighted F1 score de cada classificador, estes são apresentados na Tabela 1. Esses experimentos foram realizados para avaliar qual classificador de aprendizagem de máquina apresentaria o melhor desempenho.

Tabela 1. Resultado dos melhores experimentos.

Método	Precisão	*Macro F1*	F1 Ponderada
MLP 1 (1 cam. oculta, 500 neurônios)	84.89%	84.21%	84.99%
MLP 2 (2 cam. ocultas, 800 e 500 neurônios)	87.62%	87.44%	87.70%
Regressão Logística	84.89%	82.75%	84.75%
Árvore de Decisão	71.86%	63.95%	71.98%
Random forest	80.23%	76.09%	79.53%
Extra trees	78.46%	76.71%	78.03%
K-Nearest Neighbors	85.05%	83.93%	85.20%

Finalizada a avaliação de classificadores, definiu-se o **Experimento de reconhecimento de entidades**, utilizando o modelo BERT. Não foi possível utilizar o DGO para realizar os experimentos, pois há necessidade de anotar os dados, portanto optou-se avaliar o modelo BERT utilizando o *dataset* Drug-Drug Interaction (DDI) [Segura Bedmar et al. 2013]. O DDI é semanticamente anotado de documentos que descrevem as interações medicamentosas do banco de dados DrugBank e resumos do MedLine sobre o assunto. *Dataset* disponibilizado no DDIExtraction 2013 para execução de uma tarefa desafio, com corpus de 1.017 textos (784 textos do DrugBank e 233 resumos do MedLine), anotado manualmente com um total de 18.491 substâncias farmacológicas e 5.021 interações medicamentosas. O *dataset* é distribuído em documentos XML seguindo o formato unificado, o qual dividiu-se a fim de construir os conjuntos de dados para o treinamento e teste. Foram selecionados aleatoriamente 77% dos documentos para o conjunto de dados de treinamento e os restantes (142 textos do DrugBank e 91 resumos do MedLine) foram usados para o conjunto de dados de teste. O conjunto de **dados de treinamento** é o mesmo para ambas, pois contém entidades e anotações DDI. O conjunto de **dados de teste** foi formado descartando documentos que continham anotações DDI. Os documentos restantes foram usados para criar o conjunto de dados de teste.

Para as tarefas REN e ER adotou-se o modelo *Transformer* no domínio de medicamentos–futuramente, no domínio de textos oncológicos. Para a definição da **Arquitetura do experimento**, utilizou-se como base a documentação presente no Hug-

gingface[2], do qual foi trabalhado com o modelo BertForTokenClassification, proposto em [Devlin et al. 2018]. O modelo possui um bom suporte em bibliotecas, tais como um *tokenizer* Python (chamado "lento") e um *tokenizer* "rápido", sendo compatível também com Jax (via Flax), PyTorch e TensorFlow. Neste experimento, o BERT foi treinado para efetuar o reconhecimento das entidades, assim como a extração de características que representarão o texto utilizando o DDI. Foram experimentados diferentes modelos a fim de identificar o melhor resultado para tarefas REN e ER com a Bi-LSTM. Os modelos BERT selecionados foram: "BERT-base-multilingual-cased" [de Souza et al. 2021], "BioBERT-PT" [de Souza et al. 2021] [Ji et al. 2020] e "BERTimbau" [Lopes et al. 2021] [Souza et al. 2020]. Alguns exemplos em relação ao *dataset* são destacados a seguir: O texto base é "*Prolonged recovery time may occur if barbiturates and/or narcotics are used concurrently with ketamine.*" e possui como entidades: Nome: "*barbiturates*", Tipo: "grupo"; Nome: "*narcotics*", Tipo: "grupo"; e Nome: "*ketamine*", Tipo: "droga". As **Relações** são: "Efeito" ("*barbiturates*", "*ketamine*"); "Efeito" ("*narcotics*", "*ketamine*"); e "Vazio" ("*barbiturates*", "*narcotics*"). Através do uso do BERT, é possível identificar as entidades ao classificar cada palavra ("comum", significa que é um termo comum, não uma entidade).

4.1. Avaliação dos resultados

Na avaliação, inicialmente analisou-se um experimento que considera cada palavra individualmente no reconhecimento de entidades, seguido de um segundo resultado que foi implementar um algoritmo que decide se quer uni-las ou não. Ambos são descritos a seguir. No caso do reconhecimento de palavras individuais, para o cálculo do desempenho desse resultado ainda é considerada individualmente cada palavra. Por exemplo, caso haja a entidade "dipirona sódica"e o experimento identificar como "remédio"e "não-entidade", considera-se como 1 acerto e 1 erro, enquanto, no melhor cenário, deveria considerar como 1 erro ou 1 acerto parcial. Cada exemplo trabalhado no *dataset* é composto por um conjunto de quatro pares chave-valor. As chaves são: *Tokens* são os termos de cada texto; *True_types* são os tipos corretos das entidades; *Predicted_types* são as predições do BERT, juntamente com a probabilidade que ele deu para cada classe; *Fold* é o *fold* do exemplo.

Utilizou-se o procedimento de validação cruzada, onde separou-se o *dataset* em 5 subconjuntos, treinando 5 modelos diferentes. Para cada, foi utilizado 1 subconjunto como teste, 1 subconjunto como validação, e 3 subconjuntos como treinamento. Os resultados são apenas para os conjuntos de teste. A tabela 2 mostra o desempenho do resultado. A coluna Suporte apresenta quantos exemplos de cada classe estão presentes no *dataset*.

Tabela 2. Desempenho por palavras individuais.

	Precisão	***Recall***	***F1 score***	**Suporte**
Brand	94,65%	95,80%	95,22%	1571
Drug	95,34%	94,38%	94,86%	10276
Drug_n	74,60%	73,75%	74,17%	1063
Group	89,11%	92,37%	90,71%	6334
Outra	99,40%	99,30%	99,35%	137681

[2]www.huggingface.co/transformers

Accuracy			98,49%	156925
Macro avg	90,62%	91,12%	90,86%	156925
Weighted avg	98,50%	98,49%	98,49%	156925

Conforme visto na tabela 2, o experimento apresenta bons resultados, chegando a 98,49% de *Accuracy* nos *Tokens*. No caso de reconhecimento de palavras em grupo, para obter esses resultados, trabalhou-se no problema de encontrar os limites de cada entidade dentro dos textos. O formato é similar ao anterior, contendo duas principais alterações. A primeira alteração consiste em adicionar uma classe especial para detectar o início de um tipo de entidade. Dessa forma, é possível encontrar uma quebra quando duas entidades diferentes estiverem separadas só por um espaço em branco. Sendo assim, nas métricas de desempenho por "*token*", também aparecerá classes com um prefixo "inicio"(Tabela 3). Na segunda alteração o procedimento dedica-se à detecção das entidades por completo.

Tabela 3. Desempenho de Tokens e Entidades dos experimentos.

	Precisão	***Recall***	***F1 score***	**Suporte**
Brand - Tokens	96,06%	92,78%	94,39%	1551
Drug - Tokens	95,63%	93,43%	94,52%	9970
Drug_n - Tokens	75,69%	65,58%	70,27%	921
Group - Tokens	89,44%	88,00%	88,71%	5898
Inicio_*brand* - Tokens	21,52%	85,00%	34,34%	20
Inicio_*drug* - Tokens	47,49%	80,39%	59,71%	306
Inicio_ *drug_n* - Tokens	58,00%	81,69%	67,84%	142
Inicio_*group* - Tokens	53,79%	87,84%	66,72%	436
Outra - Tokens	99,39%	99,32%	99,35%	137681
Brand - Entidades	98,61%	97,87%	98,24%	2254
Drug - Entidades	98,78%	97,74%	98,25%	15026
Drug_n - Entidades	90,75%	70,38%	79,28%	655
Group - Entidades	97,76%	95,19%	96,46%	5323
Outra - Entidades	0	0	0	0
Accuracy - Tokens			98,17%	156925
Macro avg - Tokens	70,78%	86,00%	75,10%	156925
Weighted avg - Tokens	98,33%	98,17%	98,22%	156925
Accuracy - Entidades			98,17%	156925
Macro avg - Entidades	70,78%	86,00%	75,10%	156925
Weighted avg - Entidades	98,33%	98,17%	98,22%	156925

Mesmo com a aplicação em grupos de palavras, o experimento continuou apresentando um bom resultado, exibindo uma *accuracy* de 98,17%. Tendo como base a frase "*Other Drugs: Based on the results of drug interaction studies, no dosage adjustment is recommended when SUSTIVA (efavirenz) is given with the following: aluminum/magnesium hydroxide antacids, azithromycin, cetirizine, famotidine, fluconazole, lamivudine, lorazepam, nelfinavir, paroxetine, and zidovudine.*", cada *Token* possui uma

Tru_types, que representa a classe correta daquela palavra e a classe que o modelo reconheceu *(predicted_types)*. Os resultados obtidos são para o *Token* Drugs e *Tru_types* outra são *predicted_types* outra com 99,95% e para o *Token* adjustment e *Tru_types* outra o resultado obtido de 99,97% para outra. Os seguintes *Tokens* foram: *Token* SUSTIVA, *Tru_types* brand, *predicted_types* brand - 74,81%; *Token* efavirenz, *Tru_types* drug, *predicted_types* drug - 63,64%; *Token* aluminum, *Tru_types* group, *predicted_types* drug - 64,70%; *Token* antacids, *Tru_types* group, *predicted_types* group - 68,51%; *Token* azithromycin, *Tru_types* drug, *predicted_types* drug - 79,10%.

Esses resultados serão treinados na Bi-LSTM, para que possam ser aplicados experimentos de ER, a fim de extrair pares de relações entre as entidades. No **Experimento de extração de informações**, aplicou-se uma Bi-LSTM com objetivo de identificar pares de entidades (ou triplas) e seus tipos, classificando as relações existentes. A Bi-LSTM será executada diversas vezes separadamente para cada par de entidades, trio, quadra, etc. *Tokens* especiais indicarão o início e o fim das entidades (que serão reconhecidas pelo BERT). Será realizada uma implementação PyTorch da API Hugging Face, adicionando no topo dos modelos BERT um classificador de nível de *Token*. Serão utilizados os dados do DDI, principalmente relacionados ao experimento anterior. No texto de exemplo, três pares serão classificados, além do único trio. A Tabela 4 ilustra esse exemplo.

Tabela 4. Exemplos de relação

Texto de Entrada	Relação
Prolonged recovery time may occur if «I_ENT» barbiturates «F_ENT» and/or «I_ENT» narcotics «F_ENT» are used concurrently with ketamine.	Vazio
Prolonged recovery time may occur if «I_ENT» barbiturates «F_ENT» and/or narcotics are used concurrently with «I_ENT» ketamine «F_ENT».	Efeito
Prolonged recovery time may occur if «I_ENT» barbiturates «F_ENT» and/or «I_ENT» narcotics «F_ENT» are used concurrently with ketamine.	Efeito
Prolonged recovery time may occur if «I_ENT» barbiturates «F_ENT» and/or «I_ENT» narcotics «F_ENT» are used concurrently with «I_ENT» ketamine «F_ENT».	Vazio

5. Conclusão

Este estudo apresentou a pesquisa e definição de um novo modelo para extração de relações em dados não estruturados. Conduziu-se uma revisão bibliográfica dos formalismos e trabalhos relacionados, bem como a definição da estrutura e contexto geral do modelo. O estudo de caso e a experimentação proposta permitiu a elaboração de experimentos preliminares com arquiteturas *Transformers* e modelos de linguagem BERT. Este modelo será implementado como **caso de estudo** no SGO e seus resultados serão usados para auxílio na área da saúde oncológica. O modelo será avaliado por especialistas na área da saúde e também a nível computacional. A parceria com a Interprocess (empresa responsável por fornecedor os dados para a pesquisa sob o processo número 159593/2019-0 do CNPQ), junto ao programa Doutorado Acadêmico em Inovação , além da oportunidade de gerar inovação no setor empresarial, possibilita uma aproximação em relação ao conhecimento técnico de especialistas em saúde, fazendo com que a aproximação com especialistas da área da saúde e da área de informática médica torne-se um diferencial para

essa pesquisa. Os **resultados preliminares** evidenciaram o potencial e a importância desta pesquisa para a área da saúde, ao identificar as possibilidades a serem trabalhadas para uma extração eficaz de relações e aspectos a avaliar sobre a forma de estruturar esses dados recebidos em LN. **Como trabalhos futuros**, será realizado a criação e anotação do *Dataset* com os dados do SGO, com apoio de especialistas da saúde. A segunda é a realização de novos experimentos de Reconhecimento de entidades utilizando o DGO. Para isso será treinada uma rede Bi-LSTM para extração de relações utilizando tanto os dados do dataset já utilizado (DDI) como do DGO. Serão realizados experimentos comparativos com o *Fine-tuning* das versões atualmente utilizadas do BERT e também com a geração de uma versão customizada de BERT para os dados médicos do estudo de caso com dados oncológicos em Português. Na terceira etapa uma ontologia será populada com dados estruturados resultantes dos experimentos do trabalho. Por fim, o modelo será implementado no SGO para avançar em aspectos do estudo de caso, afim de serem realizados testes e validações. Após estas etapas, o modelo será publicado e disponível.

6. Agradecimentos

Os autores agradecem à CAPES (Código Financeiro 001) e ao CNPQ (Bolsa nº 159593/2019-0 e 309537/2020-7) por apoiar este trabalho.

Referências

Adlung, L. et al. (2021). Machine learning in clinical decision making. *Med*.

Alemzadeh, H. and Devarakonda, M. (2017). An nlp-based cognitive system for disease status identification in electronic health records. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 89–92. IEEE.

Bucur, A. et al. (2013). Clinical decision support framework for validation of multiscale models and personalization of treatment in oncology. In *13th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–4. IEEE.

Christopoulou, F. et al. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Cuocolo, R. et al. (2020). Machine learning in oncology: a clinical appraisal. *Cancer letters*, 481:55–62.

de Souza, J. V. A. et al. (2021). A multilabel approach to portuguese clinical named entity recognition. *Journal of Health Informatics*, 12.

Devlin, J. et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhole, G. and Uke, N. (2014). Nlp based retrieval of medical information for diagnosis of human diseases. *Int J Renew Energy Technol*, 3(10):243e8.

Ford, E. et al. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015.

Gonzalez-Hernandez, G. et al. (2017). Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(1):214.

Ji, Z. et al. (2020). Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Koleck, T. A. et al. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379.

Kreimeyer, K. et al. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Lee, H. et al. (2016). Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 957–960.

Li, Y. et al. (2020). Behrt: transformer for electronic health records. 10(1):1–12.

Lopes, É. et al. (2021). Exploring bert for aspect extraction in portuguese language. In *The International FLAIRS Conference Proceedings*, volume 34.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.

Morgan, S. et al. (2021). Wlv-rit at germeval 2021: Multitask learning with transformers to detect toxic, engaging, and fact-claiming comments. *arXiv preprint arXiv:2108.00057*.

Ngiam, K. Y. and Khor, W. (2019). Big data and machine learning algorithms for healthcare delivery. *The Lancet Oncology*, 20(5):e262–e273.

Prisma (2021). Preferred reporting items for systematic reviews and meta-analyses. Disponível em: <http://prisma-statement.org/PRISMAStatement/Checklist.aspx>. Acessado em 08/04/2021.

Schneider, E. T. R. et al. (2020). Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Segura Bedmar, I. et al. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Souza, F. et al. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Xue, K. et al. (2019). Fine-tuning bert for joint entity and relation extraction in chinese medical text. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 892–897. IEEE.

Yang, X. et al. (2019). Madex: a system for detecting medications, adverse drug events, and their relations from clinical notes. *Drug safety*, 42(1):123–133.