Análise do uso de NLP em documentos clínicos

Gabriel C. Fernandes[1]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo – RS – Brasil

gabcastro@edu.unisinos.br

1. Introdução

O suporte à decisões clínicas (CDS[1]) para ajudar profissionais da área da saúde possui grande relevância, permitindo assim que médicos possam interpretar acontecimentos passados e presentes de forma que seja possível guiar novas decisões com mais precisão e cuidados ao paciente. Sistemas hoje baseados em CDS são construídos através do histórico de dados de pacientes, de forma que possua as características individuais de cada um e por consequência gerando assim uma base de conhecimento com o propósito de realizar recomendações para decisões necessárias.

Dados de paciente são considerados como ”dados sensíveis” e necessitam de um cuidado grande para que possam ser manejados. Além de disso, esse tipo de dado normalmente é coletado manualmente e inserido em um sistema de CDS. Porém, pode ter o suporte de forma mais efetiva quando conectado com registros eletrônicos de saúde (EHR[2]), que frequentemente podem ajudar com informações de resultados em laboratórios, ordens de farmácias, e diagnósticos estruturados e codificados.

Notoriamente, a grande maioria dos históricos de pacientes, como exames físicos, resultados de radiologias e outros tipos de análises são gerados em formato livre. Existe então uma complexidade em certos momentos de tomar decisões baseadas no passado de um paciente, devido a grande quantidade de informações em alguns casos e a necessidade de se ter tudo em um só lugar. Sistemas de suporte à decisão são importantes e facilitam a rotina de médicos, contudo, com passar do tempo é notório que o volume de dados aumentou e outras formas computacionais podem estar ajudando nesse processo. [Demner-Fushman et al. 2009].

Nos recentes anos, a inteligência artificial impactou diversos campos com técnicas de reconhecimento de fala, visão computacional e linguagem natural de processamento, dentro do campo biomédico como fora também. Em notas clínicas e alguns tipos de exames, o uso de métodos de processamento de linguagem natural (NLP[3]) pode ser adicionado sem impacto e de forma que ainda permita a extração de informações juntamente com diversos tipos de regras de decisões. Segundo [Demner-Fushman et al. 2009], NLP pode aumentar a qualidade de CDS e também representar conhecimentos clínicos e intervenções de formatos padronizados. Além disso, muitos trabalhos vem buscando sumarizar as possibilidades e avanços no uso de NLP, não somente à CDS, mas também ao uso diretamente em EHR. [Li et al. 2021].

[1]CDS: Clinical Decision Support
[2]EHR: Eletronic Health Records
[3]NLP: Natural Language Processing

Neste trabalho, buscou-se analisar a literatura relacionada ao uso de métodos de linguagem natural diretamente para o contexto de documentos clínicos. Uma breve descrição dos conceitos envolvidos é detalhado na seção seguinte, assim como alguns trabalhos relacionados ao tema, considerados estado da arte, na seção 3. Na metodologia, é descrito sobre alguns experimentos com uma biblioteca de código aberto, criada diretamente para uso em textos clínicos. E também as possibilidades e dificuldades para a implementação e análise. Por fim, na conclusão é discutido sobre os resultados, retomando brevemente os pontos chaves da pesquisa realizada, juntamente com a visão sobre o tema investigado.

2. Conceitos gerais

Nessa seção, é apresentado uma visão geral de conceitos importantes sobre processamento de linguagem natural. Na sequência, é descrito brevemente sobre documentos clínicos, como registros eletrônicos de saúde.

2.1. Processamento de Linguagem Natural

A linguagem natural possui uma vasta gama de regras e palavras, com diferentes significados, formas de escrever, ambiguidade, etc. Pela natureza irrestrita e ambiguidade, dois problemas surgiram ao usar abordagens de análise padrão que dependiam puramente de regras simbólicas e artesanais: regras numerosas e incontroláveis, com análises ambíguas (diversas interpretações de uma única sequência de palavras); regras escritas a mão lidam com o dialeto falado e muitas vezes não sendo totalmente gramatical, e em contextos médicos, ainda pode-se destacar a fala e escrita muitas vezes considerada ”altamente telegráfica”, pela omissão de palavras em notas intra-hospitalares.

As dificuldades apresentadas fizeram com que houvesse uma reorientação onde aproximações simples e robustas substituíram análises profundas, avaliações tornaram-se mais rigorosas, métodos de Machine Learning (ML) usando probabilidades tornaram-se mais proeminentes, e por fim, uma grande quantidade de anotações (*corpora*) foram empregadas para a realização dos treinos em ML. [Nadkarni et al. 2011].

2.1.1. NLP Pipeline

Tipicamente, um fluxo de processo em NLP pode ser composto pelos seguintes componentes (Figura 1):

- Detector de sentenças: Um modelo para determinar do ponto de vista estatístico se um caractere de pontuação delimita ou não o final de uma frase. O comprimento da frase é baseado no número de caracteres disponíveis entre dois sinais de pontuação. O primeiro caractere que não é espaço em branco é considerado o início da frase, e o último caractere que não é espaço em branco define o final. A detecção de sentença é o primeiro passo na execução da NLP antes de fazer a tokenização do texto. Contudo, em contextos médicos, a abreviações e títulos (‘m.g.’, ‘Dr.’) complicam esta tarefa, assim como itens em uma lista ou enunciados padronizados (por exemplo, ‘MI [x], SOB[]’).
- Tokenizer: Responsável por dividir uma sequência de caracteres em diferentes tokens. É feito sobre sentenças, em que os tokens são as informações mínimas a

Figure 1. Fluxo em NLP. [Menasalvas et al. 2018].

serem utilizadas. Esses pequenos pedaços são normalmente palavras, símbolos de pontuação, números, etc. Tokens frequentemente contém caracteres tipicamente usados como limitadores, como hífen ou barra para frente, contudo em textos biomédicos, é comum ver o uso como parte de uma palavra ('10 mg/dia', 'N-acetylcysteine').

- Part-of-Speech (PoS): Conhecido como tagger, o PoS é responsável por atribuir tags aos diferentes tokens disponíveis para sua categoria ou tipo de palavra correspondente, dependendo dessa função do ponto de vista sintático. Além disso, dependendo da linguagem, há complicadores como gerúndio, em inglês (por exemplo, verbos terminados com 'ing' que são usados com substantivos).
- Chunker (também conhecido como Shallow Parsing): Este processo consiste em dividir um texto em partes sintaticamente correlacionadas de palavras, como sintagmas nominais ou frases verbais, mas sem especificar sua estrutura interna ou papel na frase.
- Parser: Permite a construção da árvore com o significado sintático de uma frase, dividindo-a em sujeito, verbo e objetos, bem como definindo a relação entre eles.
- Reconhecimento de Entidade Nomeada (NER[4]): É um dos últimos processos e define os tokens ou frases com seus significados semânticos ou categorias correspondentes. Por exemplo, pessoa, localização, doença, genes, ou medicação. Uma tarefa comum realizada por NER é o mapeamento de entidades nomeadas para conceitos do vocabulário. Essa tarefa geralmente aproveita a análise superficial para entidades candidatas (por exemplo, a frase nominal 'sensibilidade no peito'); no entanto, às vezes o conceito é dividido em várias frases (por exemplo, 'parede torácica mostra leve sensibilidade à pressão'). [Nadkarni et al. 2011].

Ainda existe a possibilidade de utilização da lematização, que realiza a conversão de uma palavra para a sua 'raiz', removendo sufixos. Por fim, segundo [Nadkarni et al. 2011], há muitos problemas que fazem com que NER seja desafiador: (1) variação das frases/palavras; (2) derivação: um sufixo pode transformar uma palavra em outra; (3) inflexão; (4) sinônimo; (5) homógrafos.

2.1.2. Outros métodos

Além do uso das técnicas demonstradas anteriormente, para extração e entendimento de textos clínicos, muitos estudos atualmente buscam usar juntamente Deep Learning (DL), devido as grandes possibilidades na área. Resumidamente, pode-se citar algumas arquiteturas, como: Autoencoders; Convolutional Neural Networks; Recurrent Neural Networks; Sequence-to-sequence (seq2seq); Word Embeddings; e Transfer Learning. Todas essas técnicas, juntamente com um processo de extração de informações, normalização

[4]NER: Named Entity Recognition

do conjunto de dados, acaba proporcionando uma arquitetura computacional mais robusta e aplicável à diversos contextos.

2.2. Visão geral sobre registros eletrônicos de saúde

Registros eletrônicos de saúde abrem oportunidades para melhorar o cuidado à pacientes, incorporar medidas de desempenho na prática clínica e melhorar a identificação e o recrutamento de pacientes e profissionais de saúde elegíveis em pesquisas clínicas. Os EHRs representam dados longitudinais (em formato eletrônico) que são coletados durante a prestação de cuidados de saúde rotineiros. EHRs contêm dados demográficos, estatísticas vitais, administrativos, de reivindicação (médicos e farmacêuticos) clínicos e centrados no paciente (por exemplo, provenientes de instrumentos de qualidade de vida relacionados à saúde, dispositivos de monitoramento domiciliar e avaliações de fragilidade ou cuidador). O EHR pode refletir componentes únicos de cuidados (por exemplo, atenção primária, departamento de emergência e unidade de terapia intensiva) ou dados de um sistema integrado em todo o hospital ou inter-hospitalar. [Cowie et al. 2017].

Ainda segundo [Cowie et al. 2017], os desafios para o uso de EHRs em ensaios clínicos foram identificados, relacionados à qualidade e validação de dados, captura completa de dados, heterogeneidade entre sistemas e desenvolvimento de um conhecimento de trabalho entre sistemas. O sucesso desses esforços tem que estar ligado a um planejamento cuidadoso por um grupo de várias partes interessadas e comprometida com a privacidade do paciente, segurança de dados, governança justa, infraestrutura de dados robusta e ciência de qualidade desde o início.

Por isso, é sempre necessário entender o contexto dos registros, e cuidando sempre para prezar um uso de forma restrita dos dados. Na seção seguinte é apresentado mais sobre o uso de EHRs, e as possibilidades de uso desse formato de dados juntamente com técnicas de NLP.

3. Estado da arte

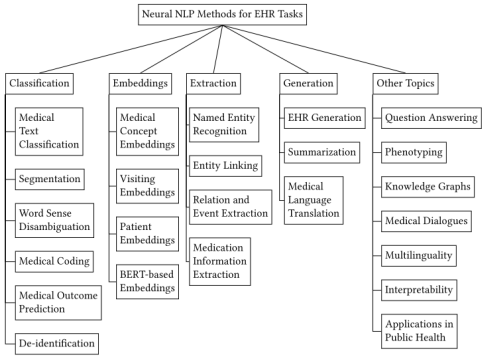
Nessa seção é apresentado trabalhos existentes e relevantes em relação à utilização de NLP aplicado à campos da biomedicina, como registros eletrônicos. A técnica de processamento de linguagem natural permite realizar correlações e mapeamento de informações textuais, permitindo assim análises e criação de ferramentas, como é mostrado abaixo.

3.1. Registros eletrônicos de saúde

Registros eletrônicos de saúde (EHR) são uma coleção de eventos e observações sobre a saúde de um paciente. Os dados contidos em EHRs podem estar estruturados ou não-estruturados, onde: EHR estruturados consiste de fontes homogenias como diagnósticos, medicações ou exames laboratoriais; já dados não-estruturados consistem de textos em formato livre disponibilizados por fornecedores, como notas clínicas e resumos de alta. Segundo [Li et al. 2021], dados não-estruturados representam aproximadamente 80% do total de registros eletrônicos, e são difíceis de serem usados em uma segunda instância.

Ainda, segundo os autores, a área possui diversos desafios e dificuldades, como: privacidade, pela sensibilidade em que os dados se encontram, e diversos órgãos com leis regulatórias; falta de anotações, pela grande escala de dados não-estruturados há a necessidade de um profissional que faça tal tarefa, contudo isso é exaustivo e difícil de

Figure 2. Tarefas divididas por métodos. [Li et al. 2021].



mensurar, o que torna um desafio para uso em algoritmos de ML supervisionados; interpretabilidade, é uma diferença nesse campo pois o uso de DL, por exemplo, ainda é considerado uma caixa-preta, o que faz com que haja a necessidade de realizar investigações sobre o uso de Explicabilidade para possibilitar uma maior transparência.

Recentemente, não somente dados estruturados mas dados em formato livre se beneficiaram das técnicas de processamento de linguagem natural, onde é possível extrair diversas informações e ter insights que permitem melhorar a vida de pacientes. Para [Li et al. 2021], os métodos atualmente aplicados em EHR, podem ser divididos como na Figura 2.

Diversas pesquisas focaram na variedade de tópicos sobre o uso DL dentro de campos como informática para saúde e bioinformática, de forma a incluir EHRs. Em [Miotto et al. 2018] é resumido diversos métodos que buscam assistir medicamente um paciente de uma forma mais adequada, rápida ou com maior precisão. Para isso, foi abordado diversos métodos e suas aplicações em imagens clínicas, EHRs, genômica e domínios mobile. Em uma pesquisa mais antiga, [Featherly 2005] busca revisar e mostrar as técnicas de DL usadas para aplicar em EHRs, juntamente com aplicações e frameworks projetados.

Outra variação da utilização de EHRs foi apresentada por [Menasalvas et al. 2018], na qual lidaram com a narrativa clínica de pré-processamento usando NLP, reconhecimento de entidade nomeada, enriquecimento semântico e por fim, integração de resultados. Para isso foi usado relatórios e anotações de pacientes que sofriam da doença Alzheimer, assim como de pacientes com câncer de pulmão (tratamento, antecedentes e outras informações), com o intuito de ajudar médicos a terem um gerenciamento melhor das doenças.

3.2. NLP aplicado à documentos clínicos

Uma larga quantidade de conhecimentos sobre biomedicina e comunicações clínicas são codificados de forma livre, seja em notas clínicas ou literatura. Nos últimos anos, a comunidade de pesquisadores conseguiu avançar de forma substancial no uso e melhoramento das técnicas de processamento de linguagem natural. Através desses avanços, análises e ferramentas foram possíveis de serem criadas.

Em [Kocaman and Talby 2021], os autores apresentaram uma abordagem de

mineração de textos clínicos, através de três pontos chaves: (1) reconhecimento de aproximadamente 100 diferentes tipos de entidades (anatomia, fatores de riscos, etc); (2) pipeline para o processamento de texto; (3) uso de modelos DL com integração em outro modelo estado da arte já pré-treinado, com diversas entidades. Dessa forma foi possível analisar a frequência de desordens, sintomas, e sinais vitais - através de dados relacionados ao COVID-19[5]. Por fim, os autores descrevem a utilização da biblioteca Spark NLP e o uso juntamente com GPU's e pipelines para melhorar a performance.

Já [Qi et al. 2020] apresentou uma ferramenta para diversas linguagens humanas usando NLP, denominada de *Stanza*. A ferramenta foi focada em atender os pipelines de um processamento de linguagem natural usando analise de texto, como *tokenização*, expansão de token com palavras multiplas, lematização, *part-of-speech*, características morfológicas, análise de pendência e reconhecimento de entidade nomeada. Em conjunto [Zhang et al. 2021] utilizou a ferramenta Stanza para a criação de um pacote, usando a linguagem Python, com foco em adaptar para o domínio biomédico e textos clínicos.

Na próxima seção será abordado um pouco mais sobre a utilização de ferramentas como a Stanza, para a análise de dados clínicos, assim como os desafios de análises em dados de contexto médico.

4. Metodologia e Resultados

Para realizar uma análise sobre linguagem natural, existem diversos estudos que concretizaram em forma de bibliotecas ou toolkits, principalmente para a linguagem de programação Python. Pode-se citar bibliotecas como MIMIC-Extract[6], ScispaCy[7], medspaCy, SciFive, EHRKit, e Stanza, na qual foi escolhida para análise dentro do presente trabalho.

Stanza, é um pacote que contém ferramentas, que podem ser usadas em um pipeline para converter strings contendo texto de linguagem humana em listas de frases e palavras, para gerar formas básicas dessas palavras, part of speech e características morfológicas, para fornecer uma análise de dependência de estrutura sintática, e para reconhecer entidades nomeadas. Segundo [Zhang et al. 2021], ferramentas NLP (toolkits) que permitem o entendimento da estrutura linguística de textos clínicos e biomédicos são frequentemente usados como a primeira etapa para a construção de um sistema. Contudo, ainda deve-se ter noção que nem todas toolkits permitem fazer análises precisas, dependendo do contexto, visto que possuem um propósito mais generalista.

A biblioteca Stanza, criada por [Qi et al. 2020], demonstrou a possibilidade de alcançar o estado da arte e funcionalidades NER, além de estendida para ser capaz de compreender o domínio clínico e biomédico.

Outro contexto, para a escolha da ferramenta, é a de que Stanza possui modelos incorporados, assim como em bibliotecas de DL, onde há redes já pré treinadas e prontas para o uso em diversas pesquisas. Além disso, os modelos usados variam de contextos diversos para o aprendizado nas diversas linguagens propostas, assim como o uso de corpora sobre textos biomédicos (Tabela 1)[8].

[5]COVID-19 Open Research Dataset (CORD-19)

[6]https://github.com/MLforHealth/MIMIC_Extract

[7]https://allenai.github.io/scispacy/

[8]https://stanfordnlp.github.io/stanza/

Table 1. Pipelines de Análise Sintática Biomédica e Clínica

Category	Treebank	Package Name	New Tokenization	Source Corpora	Treebank Doc
Bio	CRAFT	craft	Yes	Full-text biomedical articles related to the Mouse Genome Informatics database; general English Web Treebank.	CRAFT homepage
	GENIA	genia	No	PubMed abstracts related to "human", "blood cells", and "transcription factors".	GENIA homepage
Clinical	MIMIC	mimic	Yes	All types of MIMIC-III clinical notes; general English Web Treebank.	MIMIC-III homepage

O treino de modelos para identificação dos conceitos abordados na seção 2 pode tornar-se complicado em determinados casos, tanto pela construção do pipeline e cobertura de diversas regras da linguagem ou do contexto aplicado, pelo volume de dados para treino e teste, e ainda a sensibilidade dos dados, como no caso de análises para problemas biomédicos. Esse último ponto, além de extremamente importante, é um complicador pois há quantidade de dados públicos não é grande, dependendo muitas vezes de parcerias com institutos, universidades ou hospitais.

Para então realizar uma análise do funcionamento da biblioteca, partiu-se do uso do modelo já pré-treinado disponibilizado. Stanza disponibiliza atualmente três modelos, sendo eles: CRAFT, uma coleção de textos completos de artigos relacionados ao banco de dados de informática sobre genoma de ratos; GENIA, uma coleção de resumos de artigos publicados no PubMed, relacionados à "humanos", "células sanguíneas", e "fatores de transcrição"; por último, notas provenientes do banco de dados MIMIC-III.

Além da escolha do modelo a utilizar, também foi considerado - segundo a documentação da biblioteca - o uso de um modelo NER, onde cada modelo corresponde ao suporte à uma ou mais tipo de entidades. Onde é possível escolher entre entidades[9] para análise de anatomia, genética, doenças, etc.

5. Resultados

O uso de ferramentas/bibliotecas como a apresentada nesse trabalho, mostra-se importante para o avanço de análises de NLP. E no campo da biomedicina, como foi possível notar, notas médicas ou resumos de artigos possuem um contexto muito específico, que em muitos momentos se distancia da linguagem natural e cotidiana.

O uso da biblioteca é relativamente simples, caso há o interesse de obter-se resultados quanto à tokens, lematização, ou NER. Além disso, como em diversos trabalhos e nos modelos biomédicos disponíveis, a linguagem utilizada é o Inglês, foi utilizado a mesma para a visualização das informações, uma vez que quando analisado a gramática, é importante o modelo estar alinhado com a linguagem dos dados.

Para diversas frases de um documento corpora, foi aplicado análise de tokens e qual NER era identificado. Assim como o tipo de entidade, relacionada ao modele NER usado. Por último, a gramática (part-of-speech).

Como exemplo, a frase *"respiratory ¿ pt remains intubated and vented on*

[9]https://stanfordnlp.github.io/stanza/available_biomed_models.html

Figure 3. Tipos de entidades.

```
> entity: intubated     type: TREATMENT
> entity: vented        type: TREATMENT
> entity: psv8  type: TREATMENT
> entity: peep5 type: TREATMENT
> entity: o2    type: TREATMENT
> entity: abg   type: TEST
> entity: frequent suctioning   type: TREATMENT
> entity: thick, white secretions       type: PROBLEM
> entity: trach type: TREATMENT
```

Figure 4. Tokens e BIOES NER tags.

```
> token: -> ner: O
> token: pt ner: O
> token: remains    ner: O
> token: intubated  ner: S-TREATMENT
> token: and    ner: O
> token: vented ner: S-TREATMENT
> token: on ner: O
> token: psv8   ner: S-TREATMENT
> token: /  ner: O
> token: peep5  ner: S-TREATMENT
> token: and    ner: O
> token: o2 ner: S-TREATMENT
> token: 35 ner: O
> token: %. ner: O
> token: srr    ner: O
> token: 20s    ner: O
> token: w/tv   ner: O
> token: ~700-800cc.    ner: O
> token: most   ner: O
> token: current    ner: O
> token: abg    ner: S-TEST
> token: on ner: O
> token: above  ner: O
> token: settings   ner: O
> token: :  ner: O
> token: 7.41/35/121/23/-1  ner: O
> token: .  ner: O
> token: pt ner: O
> token: continues  ner: O
> token: to ner: O
> token: require    ner: O
> token: frequent   ner: B-TREATMENT
> token: suctioning ner: E-TREATMENT
> token: for    ner: O
> token: thick  ner: B-PROBLEM
> token: ,  ner: I-PROBLEM
> token: white  ner: I-PROBLEM
> token: secretions ner: E-PROBLEM
> token: .  ner: O
> token: plan   ner: O
> token: for    ner: O
> token: trach  ner: S-TREATMENT
> token: in ner: O
> token: the    ner: O
> token: or ner: O
> token: sometime   ner: O
> token: later  ner: O
> token: today  ner: O
> token: .  ner: O
```

psv8/peep5 and o2 35%. srr 20"s w/tv 700-800cc. most current abg on above settings: 7.41/35/121/23/-1. pt continues to require frequent suctioning for thick, white secretions. plan for trach in the or sometime later today." apresentou os seguintes resultados (Figura 3 e Figura 4).

Os resultados apresentados na Figura 4, mostra o formato IOB (abreviação de inside, outside, begin), um formato de marcação comum para marcação de tokens em uma tarefa de fragmentação. Uma tag O indica que um token não pertence a nenhum pedaço. E os esquemas de marcação relacionados às vezes incluem "START/END", que consiste nas tags B, E, I, S ou O, onde S é usado para representar um pedaço contendo um único token. Pedaços de comprimento maior ou igual a dois sempre começam com o B e termine com a tag E.

Por último, pode-se destacar a possibilidade de visualizar a lematização e PoS, de cada palavra (Figura 5)

Figure 5. Parte do resultado da frase de exemplo.

word.text	word.lemma	word.pos
respiratory	respiratory	ADJ
->	->	PUNCT
pt	pt	NOUN
remains	remain	VERB
intubated	intubate	VERB
and	and	CCONJ
vented	vent	VERB
on	on	ADP
psv8	psv8	NOUN
/	/	PUNCT
peep5	peep5	NOUN
and	and	CCONJ
o2	o2	NOUN
35	35	NUM
%.	%.	PUNCT
srr	srr	NOUN
20s	20	NOUN
w/tv	w/tv	NOUN
~700-800cc.	~700-800cc.	NUM
most	most	ADV
current	current	ADJ
abg	abg	NOUN

6. Conclusão

O atual trabalho buscou analisar a utilização de NLP para o contexto biomédico, em anotação de documentos clínicos. Para isso foi investigado os conceitos e alguns trabalhos estado da arte. Também foi possível identificar a utilização de métodos de processamento de linguagem natural vem cada vez mais crescendo, principalmente referente à utilização de mais métodos computacionais, como Deep Learning. E em grande parte das bibliotecas hoje disponibilizadas, são descritas com a utilização de DL em seu interior.

Além dos métodos computacionais, note-se que a coleta de dados ainda é uma tarefa não trivial e com diversos impasses, pela sensibilidade que as informações podem conter. Existe a preocupação da comunidade e ainda diversos estudos de como e quais são as devidas precauções e formas de consumo desse tipo de dados. E por esses motivos, dados públicos e de fácil acesso são difíceis de encontrar, necessitando muitas vezes passar por uma avaliação de solicitação (como o caso do banco de dados MIMIC).

Por fim, conclui-se que NLP é uma área que cresceu muito e que tem muito ainda para se desenvolver, visto que a linguagem natural é vasta e com tamanha quantidade de expressões, dialetos, formas gramaticas e regras. Além disso, o uso de NLP para a biomedicina tem uma relevância grande, já que pode ajudar cada vez mais no processo diário de muitos médicos. Identificando informações através do histórico de pacientes, e aplicando tratamentos mais assertivos.

References

Cowie, M. R., Blomster, J. I., Curtis, L. H., Duclaux, S., Ford, I., Fritz, F., Goldman, S., Janmohamed, S., Kreuzer, J., Leenay, M., Michel, A., Ong, S., Pell, J. P., Southworth, M. R., Stough, W. G., Thoenes, M., Zannad, F., and Zalewski, A. (2017). Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772.

Featherly, K. (2005). Electronic health record. *Healthcare informatics : the business magazine for information and communication systems*, 22(2).

Kocaman, V. and Talby, D. (2021). Improving clinical document understanding on COVID-19 research with spark NLP. *CEUR Workshop Proceedings*, 2831.

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., and Radev, D. (2021). *Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review*, volume 37. Association for Computing Machinery.

Menasalvas, E., Rodriguez-González, A., and Gonzalo, C. (2018). Mining Electronic Health Records: Challenges and Impact. *Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018*, pages 747–754.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. pages 101–108.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., and Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899.