ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS APLICADA A POLÍTICAS PÚBLICAS

APLICAÇÃO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL (PLN) EM PROJETOS DE PESQUISA CLÍNICA SUBMETIDAS AO SISTEMA CEP/CONEP.

Trabalho de Conclusão de Curso (Relatório de Atividades) apresentado ao curso de Especialização em Ciência de Dados aplicada a Políticas Públicas da Escola Nacional de Administração Pública -ENAP, como requisito parcial para a obtenção do título de Especialista.

Nome Mateus Magon Rodolpho
Professor Orientador: Dr. Eduardo Monteiro

Brasília, 14/09/2022

1 – Introdução: Problema abordado, justificativa e objetivos

O Sistema CEP/CONEP foi instituído pela Norma Operacional nº 01/2013 do Conselho Nacional de Saúde, com o objetivo de oferecer proteção aos participantes de pesquisa no Brasil. Todas as pesquisas que envolvam seres humanos devem ser submetidas ao Sistema CEP/CONEP, para assegurar que o desenvolvimento das pesquisas se dará dentro de padrões éticos.(BRASIL, 2013).

O objetivo da aplicação é estruturar a informação inserida em forma de texto em linguagem natural no sistema PlataformaBR, de forma a tornar possível a análise de temas de interesse e a análise de tendências.

Atualmente, a dificuldade de se modelar em formato estruturado o conteúdo textual das informações inseridas no sistema informático, impede o direcionamento de ações que permitam o planejamento, o monitoramento e a avaliação das políticas de proteção aos participantes de pesquisas.

Essas dificuldades se dividem em dois problemas principais:

- A extensão dos campos textuais cadastrados na plataforma dificulta a pesquisa com termos específicos. Atualmente, as pesquisas são feitas utilizando uma consulta SQL com condição “like” nos campos desejados, o que torna as pesquisas computacionalmente muito custosas,
- Existe a dificuldade de se averiguar a relevância de certos termos nos estudos. Além disso, como as keywords presentes no sistema são ineficientes, é difícil averiguar tendências e similaridades entre os estudos.

O presente projeto, resumidamente, criou um pipeline que recebe como input um conjunto de documentos vindos de um campo textual do sistema (por exemplo, “Título da Pesquisa”), processa os dados a partir de um modelo que os limpa e infere tokens, morfologia, sintaxe, reconhece entidades e aplica regras de reconhecimento de padrões e retorna como output alguns métodos da biblioteca Spacy (tokens, lemmas, entities, etc) e persiste em banco de dados uma lista de tokens para cada documento que trazem palavras chaves que são mais relevantes. Estes tokens são persistidos em um banco de dados para serem utilizados em ferramentas de análise e visualização de dados. As possibilidades desta abordagem são apresentadas em Jupyter Notebooks e em um dashboard que permite a análise das informações persistidas no banco de dados.

A estruturação destas informações é de fundamental importância para a boa execução do ciclo de políticas públicas (*policy cicle*), que pode ser considerada como constituída de sete fases: 1) identificação do problema, 2) formação da agenda, 3) formulação de alternativas, 4) tomada de decisão, 5) implementação, 6) avaliação e 7) extinção (SECCHI *et al.*, p. 96). O próprio guia de análise ex-ante é explícito em citar a necessidade do uso de evidências confiáveis no processo de formulação, avaliação e monitoramento de políticas públicas (BRASIL; IPEA, p. 58):

> *Uma política pública, seus projetos e suas ações só se justificam diante de um problema público relevante e devidamente fundamentado. Para isso, é essencial a apresentação de dados quantitativos e estudos qualitativos para evidenciar a natureza e a dimensão do problema identificado e, quando possível, a sua evolução ao longo do tempo.*

Os dados estruturados produzidos pelo modelo podem ser ou não interoperados com outros dados estruturados da própria CONEP ou disponibilizados pelo Governo Federal. Alguns dos problemas de análise de políticas públicas que podemos vislumbrar que podem ter sua solução orientada pelas informações produzidas no presente projeto são as seguintes:

- Acompanhamento temporal dos temas relevantes inseridos na plataforma, de forma a melhor planejar, monitorar e orientar a análise ética feita pelo órgão competente;
- Identificar de forma mais confiável os projetos que contém temas e objetos sensíveis (ex. certos tipos de medicamentos), de forma a melhorar o controle e o monitoramento;
- Perceber quais temáticas possuem mais problemas com a análise ética, de forma a melhor planejar ações para treinamento de Comitês de Ética ou até mesmo criar orientações específicas aos analistas.

2 - Fundamentação Teórica:

Simplificadamente, o Processamento de Linguagem Natural consiste em um conjunto de técnicas para análise de textos através da tecnologia da informação. A “Linguagem Natural” se refere à necessidade de se distinguir o processamento da linguagem estruturada, cuja sintaxe e a semântica são restritas, tais como as linguagens de programação, da linguagem viva utilizada pelas sociedades humanas.

Os primeiros estudos sobre PLN datam da década de 40, advindos do esforço de quebra de códigos criptografados durante a Segunda Guerra Mundial. Posteriormente, os conceitos aprendidos começaram a ser utilizados nas tentativas de se fazer a tradução computadorizada de línguas naturais (JOSEPH, 2016).

Hoje várias bibliotecas, entendidas aqui como códigos prontos disponibilizados em repositórios, permitem a execução de ações comuns de processamento em linguagem natural.

Podemos ver um pipeline típico de processamento de linguagem natural abaixo:

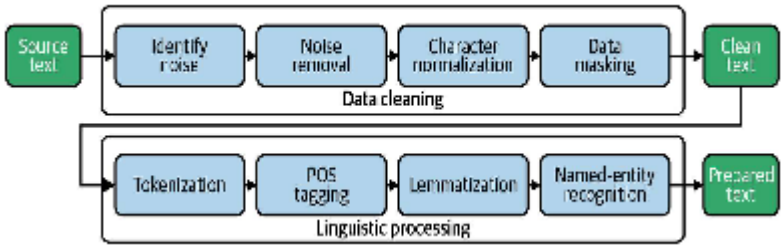

Figura 1: Um pipeline padrão para processamento de linguagem natural.

Fonte: Albrecht, J (2020)

Podemos classificar estas etapas de processamento de linguagem como segue:

- Tokenization: divisão do corpus textual em pedaços menores chamados tokens, que podem ser sentenças, palavras, letras, etc, a depender do problema que procura se resolver;
- Pós-tagging: categorização dos tokens, normalmente se utilizando a classificação morfológica das palavras;
- Lemmatization: trata-se da separação dos radicais das palavras com a utilização de modelos que agregam conforme sua semântica,
- Named-entity recognition (NER): classificação dos tokens em entidades do mundo pré-classificadas de acordo com um modelo.

Este estudo se propõe a fazer a aplicação destas técnicas em textos de pesquisa clínica não estruturados, de forma a torná-los estruturados e com maior potencial de uso.

3 – Metodologia O modelo proposto para solução do problema é construído a partir de uma solução híbrida. Parte de um modelo pré-treinado disponível na biblioteca Spacy, que se ocupa da inferência dos tokens, lemmas e entidades textuais comuns, como pessoas, organizações, localidades e países. Em virtude das peculiaridades dos textos encontrados na Plataforma Brasil, que apresenta um vocabulário que abrange termos específicos das áreas de saúde, química e de metodologia científica foram criadas uma série de regras (“rules”), capazes de identificar padrões. Estas regras foram criadas a partir de documentos que pudessem dar uma orientação por conterem uma intersecção com o campo semântico dos textos da Plataforma Brasil. Estes documentos são os seguintes:

- especialidades: elaborada a partir da RESOLUÇÃO CFM Nº 2.221/2018, com algumas mdificações para atender às necessidades do negócio,
- metodologias: lista de métodos científicos de elaboração própria,
- cid-10: foi elaborada através da tokenização, remoção de stop words e uso de técnicas estatísticas nos itens da Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde. O

uso da CID-10 se mostrou útil porque o contexto de uso das palavras (ou tokens) nos dois *corpus* se mostra mostra muito semelhante, com as palavras sendo empregadas normalmente dentro mesmo campo semântico. A fonte dos termos do CID foi extraído de banco de dados mantido pelo próprio Datasus (Ministério da Saúde) e consultado em julho de 2022.

- medicamentos: lista de substâncias obtidas através da Câmara de Regulação do Mercado de Medicamentos - CMED (https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/cmed/precos)(BRASIL; CMED, 2022).

Os motivos que levaram a opção de um modelo baseado em regras foram a falta de mão-de-obra para a anotação suficiente de dados de treinamento para alimentar uma rede neural e a falta de capacidade de processamento necessária para se treinar um modelo deste tipo. Acrescente-se que o uso de regras é uma ótima forma de construir as primeiras versões de aplicações de NLP (VAJJALA *et al*., p.749).

Todos estes processos são adicionados através do método *add* da classe NLP do spaCy e então executadas através do método *pip*" que retorna o modelo em memória e suas classes (*token, spam, ent, etc*). Por questões de praticidade os dados do modelo são persistidos em um banco de dados para uso de outros analistas.

4 – Resultados e Discussão:

O código que configura e executa o modelo encontra-se disponível em um repositório GitHub no endereço https://github.com/Dados-Abertos-Ministerio-da-Saude/plataformabr.git. Neste repositório estão também os notebooks ue demonstram os resutados obtidos. Está também em construção um dashboard para visualização dos resultados.

A construção do modelo seguiu os seguintes passos:

a) Foi feita uma "view" que processou os dados armazenados em um banco relacional no sistema em uma tabela analítica que trazia apenas os dados de interesse,
b) A partir desta tabela analítica foi feita uma análise exploratória dos dados da plataforma, com o objetivo de identificar problemas e padrões nos dados. Dentre os problemas encontrados, destaco:
    - Problemas de codificação dos caracteres, pois nem todos os registros puderam ser decodificados pelo padrão Unicode - UTF-8 (tratamento dado e comentado no código),
    - A dificuldade de se identificar termos relevantes nos textos das plataforma, mesmo depois da aplicação de técnicas como a Word-of-Bags e o TF-IDF(Term Frequency e Inverse Term Frequency

.

c) Para possibilitar a identificação de termos relevantes, foram programadas certas regras que identificam nos textos termos médicos, acadêmicos e químicos (principalmente moléculas que compõe os medicamentos). A relação dos termos também foi gerado a partir de técnicas de NLP.

d) Foi criado um arquivo "plataformabr.py", que pode ser importado em futuros códigos na forma de biblioteca que contém várias funções úteis, inclusive rodar o modelo com novos dados a partir da função *"plataformabr.modelo()"*.

Foi disponibilizado no repositório do GitHub uma série de *Jupyter Notebooks* que demonstram os problemas encontrados na análise exploratória, bem como as análises realizadas a partir do modelo pronto.

Também foi criado um *dashboard* na ferramenta Qlik Sense que é alimentado pelos dados criados pelo modelo gerado no trabalho. Seu objetivo é demonstrar as possibilidades de análise dos dados realizada após o processamento pelo modelo. Ele está disponível no endereço https://infoms.saude.gov.br/extensions/Plataforma_Brasil/Plataforma_Brasil.html.

O *dashboard* foi entregue como um aplicativo em estágio Alfa. Como se trata de uma prova de conceito, não foram feitas validações nas regras de negócios aplicadas, tampouco foram realizados os necessários testes na aplicação e confirmação das informações presentes. Estas deverão ser feitas, junto com o processo de automatização e de ETL.

No Dashboard é possível visualizar nuvens de palavras e linhas de tendência na inserção de termos no sistema por quadrimestre e ano. Os filtros permitem a pesquisa por termo geral, no campo metodologias do sistema e de substâncias. Uma alteração nos filtros permite que as informações de todos os gráficos sejam atualizadas simultâneamente, facilitando a geração de **insights**. O projeto de design segue o padrão já adotado pelo Demas/Ministério da Saúde de forma a facilitar o aproveitamento do projeto em produção (UNIVERSIDADE DE SÃO PAULO, 2022). Os códigos do modelo já estvam padronizados e as diretrizes encontram-se disponíveis em Design System SAGE (usp.br).

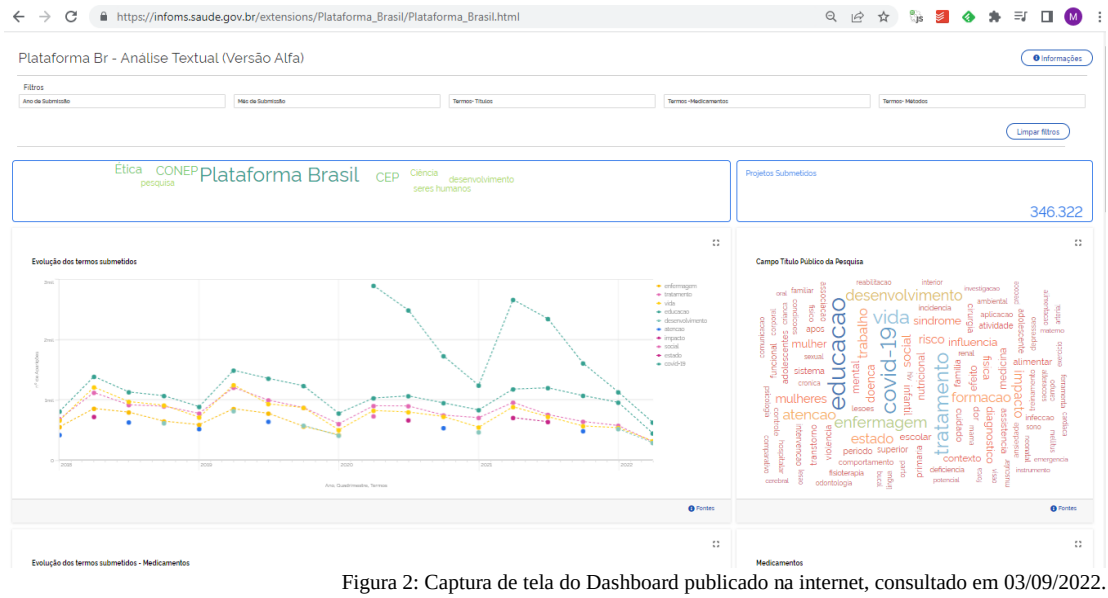

Figura 2: Captura de tela do Dashboard publicado na internet, consultado em 03/09/2022.

Todos os códigos foram disponibilizados em repositório do GitHub, da organização “Dados Abertos Ministério da Saúde” e estão disponíveis para uso e consuta do Ministério da Saúde e da ENAP. O acesso ao repositório é ontrolado, caso haja interesse de acessá-lo, entrar em contato através do e-mail mateus.rodolpho@saude.gov.br.

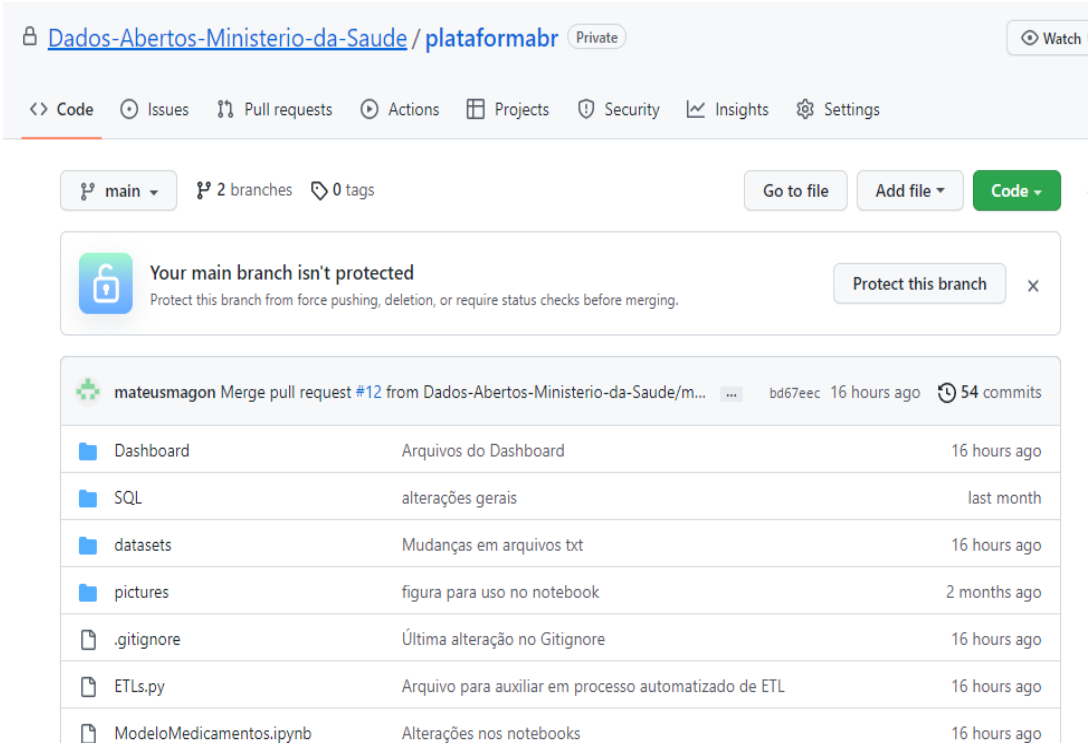

Figura 3: Captura de tela do GitHub do repositório disponiilizado, consultado em 03/09/2022.

5 – Considerações finais:

A partir do resultado do trabalho, é possível concluir que é possível utilizar as técnicas de processamento de linguagem natural nos dados armazenados no banco do Sistema PlataformaBR. O resultado da aplicação das técnicas permitiram uma simplificação da base de dados, fazendo com que a análise posterior seja computacionalmente menos custosa. É possível verificar isso a partir dos notebooks e do dashboard disponibilizado. Permitiu também que o “insight” seja mais rápido com a remoção do “ruído” presente na informação.

Entretanto, a colocação da aplicação em produção dependerá de uma melhor prospecção das regras e necessidades de negócios com os analistas e de testes que permitam a correção de possíveis erros e falhas. É importante notar que todo o projeto foi pensado para facilitar à sua incorporação pelo Ministério da Saúde, com o uso de tecnologias e padrões já utilizados pelo órgão.

Também é possível pensar para desenvolvimentos futuros, substituir a identificação de entidades baseada em regras por um baseado em aprendizado de máquina. Isso dependerá, porém, de um trabalho de

anotações em textos da plataforma para criação de dados de treino (que exigirá mão-de-obra não disponível neste momento) e de poder computacional adequado para execução dos algoritimos, o que foge do escopo deste projeto.

6 – Referências Bibliográficas:

Alguns dos códigos apresentados no repositório do GitHub foram retirados de exemplos apresentados em livros e na documentação oficial do Spacy (EXPLOSION, 2022) e das bibliotecas Python adaptados às necessidades específicas da Plataforma BR. Nesta linha, ALBRECHT, 2020, foi especialmente útil na apresentação de modelos (*blueprints*) para as análise exploratória apresentadas nos *notebooks* devidamente carregados no repostório do GitHub. Já ALTINOK , 2021, serviu a fonte que permitiu o uso do método *PhraseMatcher* do Spacy para o uso da lista CMED de medicamentos. Em ambos os casos, os códigos foram sendo adaptados de acordo com as necessidades dos problemas.

ALBRECHT, J; RAMACHANDRAN, S; WINKLER C. Blueprints for Text Analytics Using Python. O'Reilly Media, Inc. Edição do Kindle, 2020.

ALTINOK, Duygu. Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem. Packt Publishing. Edição do Kindle, 2021.

BRASIL. Conselho Nacional de Saúde. Aprova as normas regulamentadoras de Pesquisa em Seres Humanos. Norma Operacional CNS nº 1/2013.

BRASIL. Câmara de Regulação do Mercado de Medicamentos - CMED, 2022. Listas de preços de medicamentos. Disponível em: < https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/cmed/precos >. Acesso em: 15/07/2022.

BRASIL, IPEA. Avaliação de políticas públicas: guia prático de análise *ex-ante*. Volume 1. Brasília : Ipea, 2018.

EXPLOSION. spaCy, 2022. Industrial-Strenght Natural Language Processing in Python. Disponível em: < https://spacy.io/>. Acesso em: 01/08/2022.

JOSEPH, Sethunya R, *et al*. Natural Language Processing: a Review, International Journal of Research in Engineering and Applied Sciences, vol. 6, ed. 3, 2016. Disponível em https://www.researchgate.net/profile/Sethunya-Joseph/publication/309210149_Natural_Language_Processing_A_Review/links/5805ea1f08ae03256b75d965/Natural-Language-Processing-A-Review.pdf.

SECCHI, Leonardo; DE SOUZA COELHO, Fernando; PIRES, Valdemir. Políticas Públicas (pp. 96-97). Cengage Learning. Edição do Kindle.

UNIVERSIDADE DE SÃO PAULO. Faculdade de Arquitetura e Urbanismo. Design System SAGE/MS, 2021. Disponível em: < http://infovisparasaude.fau.usp.br/ds-sage/>. Acesso em: 01/09/2022.

VAJJALA, Sowmya; Majumder, BODHISATTWA; Gupta, ANUJ; Surana, HARSHIT (2020-06-17). Practical Natural Language Processing. O'Reilly Media. Edição do Kindle, 2020.