NLP aplicada à revisão sistemática de biomarcadores sanguíneos para diagnóstico da Doença de Alzheimer

Denis Kalleb Oliveira Costa[1], **Eduardo Palhares Júnior** [2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Campus Manaus Zona Leste – Manaus, AM – Brasil

soberano.dk@gmail.com, eduardo.palharesjr@ifam.edu.br

Resumo. *O diagnóstico precoce da Doença de Alzheimer representa um dos principais desafios da neurologia contemporânea. Este trabalho propõe uma abordagem automatizada baseada em técnicas de Processamento de Linguagem Natural (PLN) para análise de literatura científica sobre biomarcadores sanguíneos. A partir da extração de mais de 6.000 artigos da base Web of Science, foram aplicados modelos pré-treinados da biblioteca spaCy, extração de palavras-chave com TF–IDF, modelagem de tópicos via LDA e métricas de similaridade vetorial. Os resultados demonstram que o modelo baseado em transformadores (spaCy TRF) apresenta maior precisão semântica em tarefas de extração e agrupamento textual, evidenciando sua superioridade para análises contextuais de larga escala.*

Palavras-chave: *Alzheimer, Biomarcadores, PLN, TF-IDF, LDA.*

Abstract. *Early diagnosis of Alzheimer's disease remains one of the major challenges in modern neurology. This study proposes an automated approach based on Natural Language Processing (NLP) techniques to analyze scientific literature on blood-based biomarkers. A corpus of over 6,000 articles retrieved from the Web of Science database was processed using pre-trained models from the spaCy library. Key methodological steps included keyword extraction via TF–IDF, topic modeling using LDA, and vector-based similarity metrics. Results indicate that the transformer-based model (spaCy TRF) yields superior semantic precision for context-sensitive tasks, highlighting its effectiveness for large-scale textual analysis.*

Keywords: *Alzheimer, Biomarkers, NLP, TF–IDF, LDA.*

1. Introdução

A Doença de Alzheimer (DA) é uma enfermidade neurodegenerativa progressiva e irreversível, sendo a principal causa de demência entre idosos no mundo [Chen et al. 2023b]. O diagnóstico precoce desempenha papel fundamental na eficácia das intervenções terapêuticas, mas os métodos tradicionais — como punção lombar e exames de neuroimagem — ainda apresentam alto custo, caráter invasivo e baixa escalabilidade em contextos populacionais.

Como alternativa, biomarcadores sanguíneos têm se destacado por oferecerem uma abordagem promissora, mais acessível e menos invasiva para o rastreio e diagnóstico precoce da DA [Chen et al. 2023b]. Paralelamente, o crescimento exponencial

da produção científica na área impõe desafios à sua sistematização, exigindo ferramentas computacionais capazes de organizar, interpretar e extrair conhecimento de forma automatizada.

Neste contexto, este trabalho propõe uma abordagem baseada em técnicas de Processamento de Linguagem Natural (PLN) para realizar uma análise automatizada de mais de 6.000 artigos científicos indexados na base Web of Science (WOS). Utilizaram-se modelos pré-treinados da biblioteca spaCy, com destaque para as versões `en_core_web_sm` e `en_core_web_trf` , sendo esta última baseada em transformadores [Honnibal and Montani 2023].

A análise textual incluiu remoção de stopwords, lematização, extração de palavras-chave com TF–IDF [Ramos 2003] e modelagem de tópicos com LDA (Latent Dirichlet Allocation) [Blei et al. 2003a]. Além disso, métricas de similaridade vetorial foram aplicadas para comparar os modelos e mapear a relação entre biomarcadores e métodos diagnósticos, como PET scan, testes cognitivos e exames de fluídos.

2. Fundamentação Teórica

Esta seção apresenta os principais conceitos relacionados ao diagnóstico precoce da Doença de Alzheimer (DA) e aos métodos computacionais empregados na análise automatizada da literatura científica. Inicialmente, aborda-se a importância clínica e os desafios associados à identificação precoce da DA. Em seguida, são descritos os principais biomarcadores sanguíneos atualmente estudados. Por fim, discutem-se os fundamentos de Processamento de Linguagem Natural (PLN), incluindo as técnicas de vetorização e modelagem de tópicos utilizadas neste trabalho.

2.1. Doença de Alzheimer e Diagnóstico Precoce

A Doença de Alzheimer (DA) é uma enfermidade neurodegenerativa progressiva, caracterizada por perda de memória, declínio cognitivo e alterações comportamentais. Estima-se que mais de 55 milhões de pessoas vivam com demência no mundo, sendo a DA responsável por aproximadamente 60 a 70% dos casos registrados [World Health Organization 2022].

O diagnóstico precoce da DA é considerado essencial para intervenções terapêuticas mais eficazes, desenvolvimento de novos tratamentos e melhoria da qualidade de vida dos pacientes. Os métodos tradicionais de diagnóstico incluem exames clínicos, testes neuropsicológicos, técnicas de imagem como a tomografia por emissão de pósitrons (PET) e a ressonância magnética, além da análise do líquor obtido por punção lombar [Chen et al. 2023b]. Apesar de sua acurácia, tais métodos são onerosos, invasivos e de difícil aplicação em larga escala.

Diante disso, biomarcadores sanguíneos têm emergido como uma alternativa promissora para a detecção precoce da DA, oferecendo vantagens em termos de acessibilidade e menor invasividade [Hampel et al. 2018]. Esse cenário impulsiona o uso de métodos computacionais, como os de Processamento de Linguagem Natural (PLN), aplicados à triagem automatizada de grandes volumes de literatura científica, com o objetivo de identificar padrões relevantes e associações entre biomarcadores e estratégias diagnósticas.

2.2. Biomarcadores Sanguíneos: NFL, Beta-amiloide, GFAP, p-Tau

Biomarcadores sanguíneos têm se consolidado como ferramentas fundamentais para a detecção precoce e o monitoramento da Doença de Alzheimer (DA). Dentre os principais biomarcadores estudados, destacam-se:

- a proteína tau fosforilada (p-Tau181, p-Tau217),
- a proteína glial fibrilar ácida (GFAP),
- o neurofilamento de cadeia leve (NfL),
- e os peptídeos beta-amiloide (Aβ40, Aβ42) [Chen et al. 2023a, Palmqvist 2020].

Os níveis plasmáticos de p-Tau apresentam forte correlação com os depósitos cerebrais de tau observados por tomografia por emissão de pósitrons (PET), sendo especialmente úteis na distinção entre DA e outras demências [Karikari 2020]. O NfL, por sua vez, é considerado um marcador inespecífico de neurodegeneração, sendo eficaz no acompanhamento da progressão da doença [Preische 2019]. Já a GFAP está associada à reatividade astrocitária, com aumento detectável em fases iniciais da DA. Por fim, os peptídeos beta-amiloide são tradicionalmente relacionados ao acúmulo extracelular de placas senis, uma das principais características neuropatológicas da DA.

Estudos recentes sugerem que a combinação desses biomarcadores pode atingir alta sensibilidade e especificidade diagnóstica, especialmente quando aliada a métodos computacionais para análise em larga escala [Thijssen 2020].

2.3. Mineração de Texto e Processamento de Linguagem Natural (PLN)

As técnicas de Mineração de Texto e Processamento de Linguagem Natural (PLN) permitem extrair informações relevantes de grandes volumes de dados não estruturados, como artigos científicos. Essas abordagens têm sido cada vez mais utilizadas em revisões sistemáticas automatizadas, especialmente na área da saúde, onde há um crescimento expressivo na produção científica [Cui 2020].

Neste trabalho, adotou-se a biblioteca *spaCy*, uma das ferramentas mais modernas e eficientes para PLN. O modelo `en_core_web_sm` é leve e rápido, sendo indicado para tarefas básicas como lematização e remoção de *stopwords*. Já o modelo `en_core_web_trf`, baseado em arquiteturas do tipo *Transformer*, oferece maior precisão em tarefas contextuais e semânticas, embora exija maior capacidade computacional [Honnibal and Montani 2023].

Para a extração de palavras-chave, utilizou-se a técnica TF–IDF (Term Frequency–Inverse Document Frequency), que avalia a relevância de um termo com base em sua frequência relativa em um conjunto de documentos [Ramos 2003]. Como complemento, aplicou-se a modelagem de tópicos com o algoritmo LDA (Latent Dirichlet Allocation), que identifica agrupamentos latentes com base na coocorrência de termos nos textos [Blei et al. 2003b].

Essas técnicas viabilizaram a extração de termos recorrentes, a visualização da evolução temporal das pesquisas e a identificação de associações entre biomarcadores e terminologias diagnósticas em publicações científicas.

3. Metodologia

Esta seção apresenta o fluxo metodológico adotado para a análise automatizada da literatura científica sobre biomarcadores sanguíneos no contexto do diagnóstico da

Doença de Alzheimer. A abordagem emprega técnicas de Processamento de Linguagem Natural (PLN) para explorar um grande volume de publicações indexadas, com foco em tarefas de vetorização semântica (TF–IDF), extração de tópicos (LDA) e geração de visualizações informativas, utilizando como base modelos de linguagem pré-treinados da biblioteca spaCy [Honnibal et al. 2020].

O pipeline geral do projeto é ilustrado na Figura 1, que organiza as etapas principais do processo, desde a coleta de dados até a comparação entre modelos. As próximas subseções detalham cada componente desse fluxo de forma sistemática.

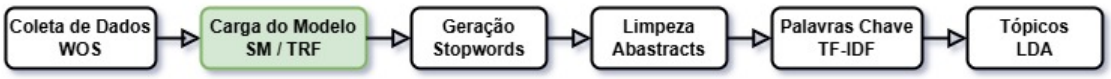

Figura 1. Fluxo metodológico do projeto

3.1. Coleta de Artigos Científicos

Os dados utilizados neste trabalho foram obtidos por meio da plataforma Web of Science (WOS), acessada via o Portal de Periódicos da CAPES. A busca foi realizada utilizando a ferramenta de pesquisa avançada, com a seguinte string:

Os metadados dos artigos foram coletados a partir da base Web of Science (WOS), acessado via o portal de periódicos da CAPES. Utilizou-se a ferramenta de busca avançada com a seguinte string:

```
("Alzheimer") AND ("p-Tau217"OR "p-Tau181"OR
"GFAP"OR "Beta-amyloid"OR "Neurofilament light"OR
"NfL"OR "biomarker") AND ("blood"OR "plasma") AND
("diagnosis"OR "early detection"OR "screening")
```

Essa estratégia resultou em um conjunto de mais de 6.900 publicações indexadas, abrangendo o período de 2018 a 2024. Os dados foram exportados em lotes no formato `.xls`, e posteriormente carregados com auxílio da biblioteca `pandas` em um único *DataFrame*, contendo as colunas `Abstract`, `Author Keywords`, `Title`, `DOI` e `Publication Year`.

Com o objetivo de enriquecer o conteúdo semântico dos textos analisados, os campos `Abstract` e `Author Keywords` foram combinados em uma única estrutura textual para cada artigo.

3.2. Pré-processamento dos Textos

Com o objetivo de preparar os textos para análise computacional, foi realizado um pré-processamento estruturado sobre os dados coletados. Inicialmente, os campos `Abstract` e `Author Keywords` foram combinados para formar uma representação textual unificada de cada artigo. Em seguida, aplicaram-se as seguintes etapas:

1. conversão de todos os caracteres para letras minúsculas;
2. remoção de pontuação, números e espaços em branco redundantes;
3. tokenização e lematização utilizando o modelo `en_core_web_sm` do *spaCy*;

4. eliminação de *stopwords*, combinando a lista padrão da biblioteca com um conjunto customizado contendo termos recorrentes do domínio, como *study*, *result*, *biomarker*, *alzheimers*, entre outros.

Adicionalmente, foi gerada uma segunda versão dos textos com o modelo `en_core_web_trf`, baseado em arquiteturas do tipo *Transformer*, utilizada posteriormente para comparações semânticas [Honnibal and Montani 2023].

3.2.1. Modelos spaCy SM e TRF

Neste trabalho, adotou-se a biblioteca `spaCy` como base para o processamento textual automatizado dos artigos científicos, devido à sua robustez, modularidade e suporte a múltiplas tarefas de PLN [Honnibal and Montani 2023].

O `spaCy` disponibiliza modelos pré-treinados otimizados para textos em inglês, os quais integram pipelines completos com componentes como tokenização, lematização, tagging gramatical (POS), análise de dependência sintática (parser) e reconhecimento de entidades nomeadas (NER). Esses modelos são treinados sobre grandes corpora públicos, como OntoNotes 5, Wikipedia e Common Crawl, tornando-os apropriados para tarefas de análise textual em larga escala [Cui and Lee 2020].

Modelo spaCy SM: O modelo `en_core_web_sm` (abreviado como SM) é uma versão leve da biblioteca `spaCy`, baseada em redes convolucionais (CNNs), projetada para oferecer alta velocidade de execução com baixo custo computacional. Ele é recomendado para tarefas básicas de Processamento de Linguagem Natural (PLN), como tokenização, lematização, análise gramatical (POS tagging), entre outras.

Neste trabalho, o modelo `SM` foi utilizado como pipeline principal de pré-processamento textual. As etapas realizadas incluíram: **tokenização**, **lemmatização**, **remoção de *stopwords*** e normalização dos textos para vetorização posterior. A escolha por esse modelo deve-se ao seu excelente equilíbrio entre desempenho e custo computacional, sendo considerado uma solução robusta para análise textual biomédica, conforme apontado na literatura especializada [González-Castro et al. 2021].

A Figura 2 ilustra a arquitetura interna do modelo, destacando os componentes ativados sequencialmente durante o processamento textual.

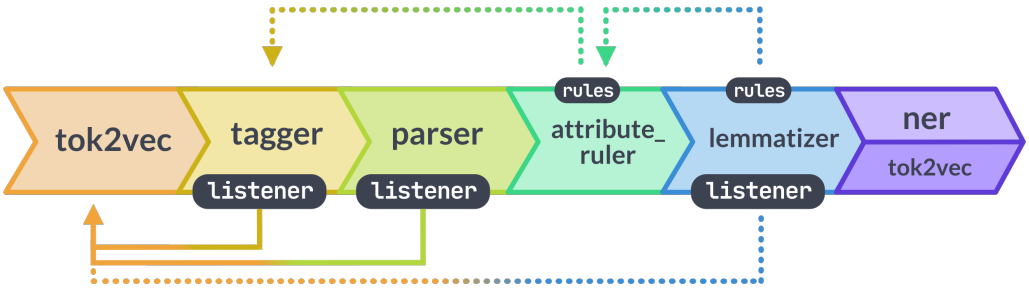

Figura 2. Pipeline padrão do modelo `en_core_web_sm` [Honnibal and Montani 2023].

Embora o pipeline inclua por padrão o componente NER (Named Entity Recognition), nesta pesquisa optou-se por não utilizá-lo diretamente. Em vez disso, adotou-se uma abordagem baseada em extração de palavras-chave com TF–IDF e modelagem de tópicos com LDA, mais adequadas à natureza estatística da revisão sistemática proposta [Cui 2020, González-Castro et al. 2021].

Modelo spaCy TRF: O modelo `en_core_web_trf` (TRF) é uma versão avançada da biblioteca `spaCy`, baseada em arquiteturas do tipo *Transformer*, como o RoBERTa. Esse tipo de modelo é especialmente indicado para tarefas que dependem de compreensão semântica de contexto, como similaridade textual, inferência lógica e reconhecimento de entidades complexas. Em contrapartida, o modelo TRF apresenta maior tempo de execução e exige mais recursos computacionais em comparação à versão leve (SM).

Neste trabalho, o modelo `TRF` foi utilizado de forma pontual, com o objetivo de comparar semanticamente diferentes versões dos textos limpos. Essa comparação foi realizada com base em métricas como a **similaridade cosseno** e a **divergência de Kullback–Leibler**, permitindo avaliar a qualidade semântica dos textos processados.

O pipeline do modelo TRF inclui os seguintes componentes ativados em sequência:

1. **Tokenizer:** Segmenta o texto em tokens, respeitando pontuações, contrações e espaços.
2. **Tagger (POS):** Atribui rótulos gramaticais aos tokens, como substantivos, verbos e adjetivos.
3. **Lemmatizer:** Reduz as palavras à sua forma canônica (ex: `studies` $\rightarrow$ `study`).
4. **Attribute Ruler:** Ajusta atributos com base em regras linguísticas predefinidas.
5. **Parser e NER:** Embora carregados por padrão, esses componentes não foram utilizados diretamente nesta pesquisa.

É importante destacar que, apesar de o componente `NER` (Named Entity Recognition) estar presente no pipeline, ele não foi empregado neste estudo. Optou-se por técnicas de vetorização baseadas em **TF–IDF** e **LDA** para a extração de palavras-chave e temas latentes, alinhando-se a estratégias estatísticas adequadas para análise de grandes corpora científicos [Cui 2020, González-Castro et al. 2021].

A adoção dessa estratégia híbrida — pré-processamento com o modelo `SM` e análise semântica pontual com o modelo `TRF` — permitiu aliar eficiência computacional e profundidade contextual, garantindo robustez na extração de conhecimento textual a partir da literatura biomédica.

3.2.2. Stopwords padrão e customizadas

As *stopwords* são termos com alta frequência de ocorrência na linguagem natural, mas com baixa relevância semântica para tarefas de análise textual. Essa categoria inclui, predominantemente, pronomes, artigos, preposições e conjunções, cuja presença tende a introduzir ruído na vetorização de textos e na modelagem semântica.

Neste trabalho, adotou-se inicialmente a lista padrão de *stopwords* da biblioteca `spaCy`, composta por termos da língua inglesa como `the`, `and`, `in`, `of`, entre outros. Complementarmente, foi construída uma lista personalizada voltada ao domínio biomédico, contendo palavras recorrentes nos resumos analisados, mas com baixo poder discriminativo, tais como `alzheimer`, `disease`, `patients`, `biomarker`, `study`, `methods` e `plasma`.

A união entre os dois conjuntos teve como objetivo aumentar a eficiência do pré-processamento, reduzindo o ruído semântico e potencializando a qualidade das etapas posteriores de vetorização, extração de tópicos e geração de palavras-chave.

Na Figura 3 podemos observar o código Python implementado para definição conjunta das listas padrão e customizada de *stopwords* utilizadas neste estudo.

```python
# Carregar arquivo já filtrado
df = pd.read_csv('D:/TCC-2024/base/artigos_filtrados.csv')

# Carregar modelo pré-treinado de linguagem inglesa
nlp = spacy.load("en_core_web_sm")

# Carregar stopwords padrão
stopwords = nlp.Defaults.stop_words

# Adicionar stopwords específicas do contexto
custom_stopwords = {
    "alzheimer", "disease", "patients", "study", "studies", "use", "method", "methods",
    "results", "data", "biomarker", "biomarkers", "plasma", "levels", "group", "groups"
}
stopwords_custom = stopwords | custom_stopwords
```

Figura 3. Definição das *stopwords* padrão e customizadas em Python

A Tabela 1 apresenta uma amostra representativa dos termos removidos do corpus, distribuídos entre as categorias padrão, contrações e termos especializados.

Tabela 1. Exemplos de palavras removidas como *stopwords*

Padrão spaCy	Contrações	Customizadas
about	's	alzheimer
above	're	biomarker
among	'm	study
after	've	methods
an	'll	patients

3.2.3. Função de limpeza textual

A função `clean_text()` constitui o núcleo do processo de pré-processamento textual, sendo responsável por normalizar e refinar os dados extraídos dos artigos científicos. Seu papel é preparar os textos para as etapas subsequentes de vetorização e análise semântica.

As operações realizadas pela função estão listadas a seguir:

1. conversão de letras maiúsculas para minúsculas;
2. remoção de números e pontuações por meio de expressões regulares;
3. tokenização com o modelo `spaCy`;
4. filtragem de tokens curtos, não alfabéticos ou presentes na lista de *stopwords*.

Foram geradas duas versões limpas do corpus, variando apenas a lista de *stopwords* aplicada: uma utilizando o conjunto padrão do `spaCy` e outra incorporando também a lista personalizada do domínio biomédico. Essa distinção permitiu avaliar o impacto da filtragem especializada na qualidade dos vetores textuais e nas etapas analíticas posteriores.

O trecho de código a seguir exemplifica a aplicação da função no `DataFrame`:

```
df["clean_default"] = df["texto_completo"].apply(lambda
x:  clean_text(x, stopwords))
df["clean_custom"] = df["texto_completo"].apply(lambda
x:  clean_text(x, stopwords_custom))
```

A coluna `texto_completo`, utilizada como entrada da função, foi construída a partir da junção dos campos `Abstract` e `Author Keywords`, com o intuito de aumentar a densidade semântica dos textos analisados.

O tempo total de execução da função `clean_text()` para todo o corpus foi de aproximadamente **19 minutos e 20 segundos**, medido com o comando `%%time` do Jupyter Notebook. A Tabela 2 apresenta um resumo das colunas resultantes desse processo.

Tabela 2. Colunas geradas no pré-processamento textual

Coluna	Descrição
`texto_completo`	Junção do campo `Abstract` com `Author Keywords`, usada como base para análise textual.
`clean_default`	Versão limpa do texto utilizando somente as *stopwords* padrão do `spaCy`.
`clean_custom`	Versão limpa utilizando *stopwords* padrão e termos personalizados do domínio biomédico.

3.3. Extração de Palavras-chave com TF–IDF

A técnica TF–IDF (Term Frequency–Inverse Document Frequency) foi aplicada para identificar os termos mais relevantes dos resumos científicos, considerando a frequência relativa das palavras em cada documento e no corpus como um todo [Ramos 2003]. Essa abordagem permite destacar termos representativos, ignorando palavras muito comuns ou irrelevantes.

Inicialmente, foi gerada uma matriz padrão de TF–IDF utilizando apenas unigramas e limitando a saída aos 50 termos com maior peso. A Figura 4 apresenta a distribuição dos termos mais relevantes identificados por essa configuração.

Observa-se que os termos mais recorrentes envolvem conceitos centrais do estudo, como `beta`, `diagnosis`, `clinical`, `cognitive` e `impairment`, reforçando a aderência do corpus ao tema proposto. No entanto, alguns termos genéricos ainda permanecem, indicando a necessidade de ajustes na parametrização.

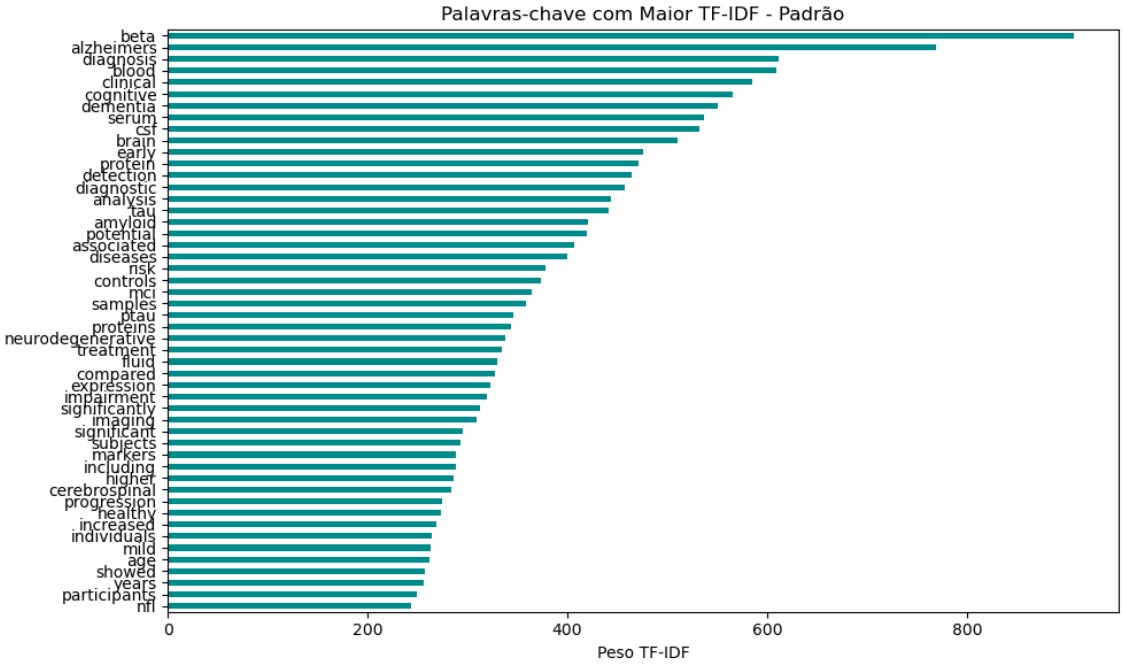

Figura 4. Distribuição de pesos dos termos mais relevantes com TF–IDF Padrão (unigramas)

Buscando melhorar o desempenho do modelo, foi gerada uma versão otimizada da matriz TF–IDF, incorporando parâmetros adicionais com foco em ganho semântico. A Figura 5 mostra os resultados obtidos com essa configuração.

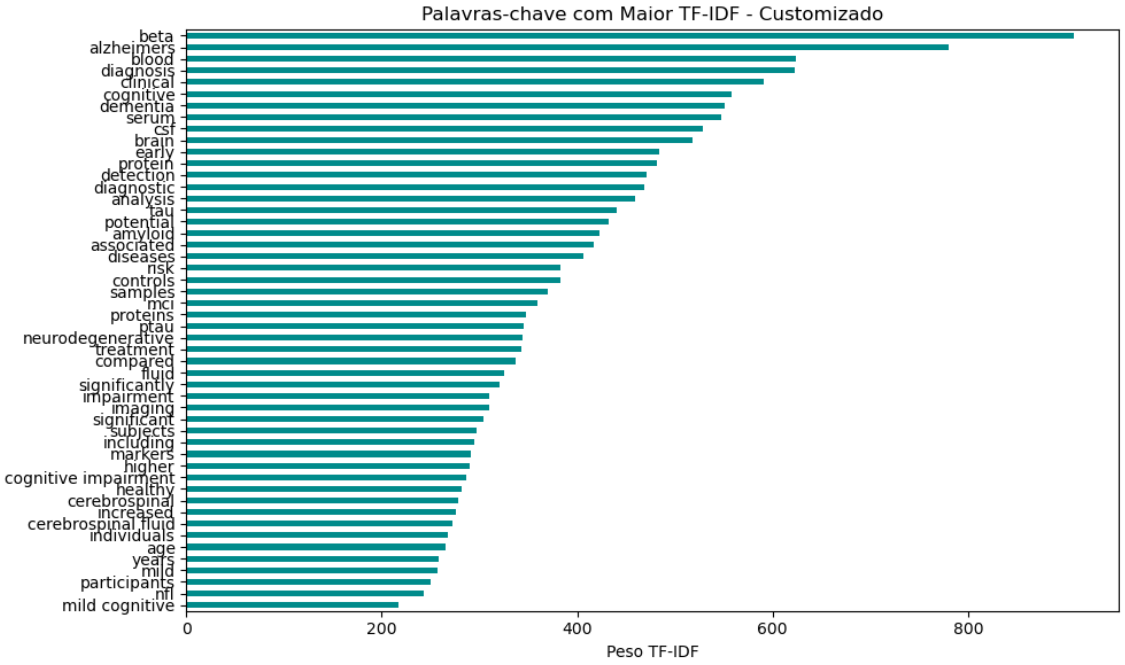

Figura 5. Distribuição de pesos dos termos mais relevantes com TF–IDF Customizado (unigramas + bigramas)

Essa versão utilizou a classe `TfidfVectorizer` da biblioteca `scikit-learn`, com os seguintes parâmetros:

- `ngram_range=(1,2)` – inclui bigramas como `cognitive impairment`;
- `stop_words='english'` – reforça a filtragem automática de palavras irrelevantes;
- `min_df=3` – remove termos raros;
- `max_df=0.85` – remove termos excessivamente frequentes.

Com essa configuração, foram extraídos termos mais específicos e compostos semanticamente ricos, como `mild cognitive`, `alzheimer disease`, `blood biomarkers` e `diagnostic value`. Isso demonstrou maior granularidade e relevância conceitual em comparação ao modelo padrão.

Na Tabela 3 podemos ver um resumo dos termos de maior peso extraídos pelas duas abordagens, permitindo comparação direta. Essa etapa foi fundamental para fortalecer as análises posteriores de modelagem de tópicos, similaridade semântica e coocorrência de termos diagnósticos.

Tabela 3. Principais palavras-chave extraídas por TF–IDF Padrão e Customizado

TF–IDF Padrão	Peso	TF–IDF Customizado	Peso
beta	920.4	beta amyloid	900.2
alzheimers	870.1	alzheimer disease	875.7
diagnosis	810.2	cognitive decline	860.3
clinical	780.6	diagnostic value	790.4
cognitive	760.2	blood biomarkers	740.6
dementia	725.4	mild cognitive	715.1
serum	710.9	early diagnosis	680.8
csf	702.3	disease progression	670.5
brain	699.1	amyloid beta	662.7
early	682.0	biomarker levels	655.9

3.4. Modelagem de Tópicos com LDA

Para identificar agrupamentos temáticos recorrentes nos artigos, foi aplicado o algoritmo Latent Dirichlet Allocation (LDA), uma técnica probabilística amplamente utilizada na análise de grandes volumes de texto [Blei et al. 2003b]. O LDA assume que cada documento é uma combinação de múltiplos tópicos latentes, e que cada tópico é definido por uma distribuição de probabilidade sobre as palavras do vocabulário.

Neste trabalho, utilizou-se a implementação `LatentDirichletAllocation` da biblioteca `scikit-learn`, com os seguintes parâmetros:

- `n_components=5` – número de tópicos a serem extraídos;
- `random_state=42` – semente para reprodutibilidade dos resultados.

A matriz de entrada foi gerada a partir da versão otimizada da vetorização TF–IDF, permitindo ao modelo capturar padrões semânticos latentes e revelar temas recorrentes na

literatura sobre biomarcadores. A Tabela 4 apresenta os cinco tópicos extraídos, com as palavras mais representativas atribuídas a cada grupo.

Tabela 4. Tópicos extraídos com LDA e suas palavras mais representativas

Tópico	Palavras-chave
1	years, associated, samples, compared, tau, diagnosis, progression, beta, potential, risk
2	diseases, proteins, amyloid, alzheimers, clinical, cerebrospinal, tau, healthy, detection, beta
3	beta, blood, fluid, treatment, diseases, cognitive, amyloid, alzheimers, impairment, brain
4	dementia, proteins, protein, serum, diseases, expression, diagnostic, blood, analysis, showed
5	dementia, subjects, alzheimers, diagnosis, nfl, neurodegenerative, controls, including, mild, compared

A modelagem com LDA permitiu identificar temas como exames de fluídos e neuroimagem (Tópico 3), neurodegeneração (Tópico 5) e progressão clínica da doença (Tópico 1). Esses resultados reforçam a consistência dos dados analisados e sua aderência à literatura sobre Alzheimer e biomarcadores.

Como perspectiva futura, técnicas baseadas em embeddings, como BERTopic ou Top2Vec, poderão oferecer maior coerência interna e granularidade temática [Pedregosa et al. 2023].

3.5. Ambiente Computacional

Após a implementação completa do pipeline de pré-processamento e análise textual, foram avaliadas as condições computacionais necessárias para execução das etapas propostas. Todos os experimentos foram realizados localmente, em um computador pessoal, sem uso de GPU ou paralelismo distribuído.

Para mensurar o desempenho computacional das etapas, foi utilizada a ferramenta `%%time` do *Jupyter Notebook*, que mede o tempo de execução de cada célula de código. Os resultados obtidos são apresentados na Tabela 5.

Tabela 5. Tempo de execução das células do notebook (via `%%time`)

Célula	Tempo de Execução
Carga dos Artigos	26.9 s
Limpeza dos Abstracts – Modelo SM	29 min 17 s
Limpeza dos Abstracts – Modelo TRF	3 h 28 min 9 s

Os resultados evidenciam a diferença significativa de custo computacional entre os modelos utilizados. O modelo `en_core_web_sm`, mais leve, foi aplicado ao corpus completo com tempo de execução inferior a 30 minutos, o que o torna apropriado para aplicações em ambientes computacionalmente limitados. Por outro lado, o modelo

`en_core_web_trf`, embora mais robusto em termos semânticos, demandou mais de 3 horas de processamento, mesmo em um corpus textual moderado.

Esses dados justificam a escolha do modelo SM como pipeline principal, relegando o uso do TRF a tarefas pontuais de comparação. De modo geral, a execução do pipeline em hardware intermediário demonstrou-se viável, reforçando a aplicabilidade prática da metodologia proposta em contextos de pesquisa com infraestrutura limitada.

4. Resultados e Discussões

Esta seção apresenta os principais resultados obtidos a partir da aplicação do pipeline de análise automatizada sobre a literatura científica. Cada conjunto de técnicas foi avaliado quanto à sua capacidade de representar e estruturar semanticamente os dados, com foco na identificação de padrões temáticos, palavras-chave relevantes e similaridade entre documentos.

As análises estão organizadas em blocos temáticos, incluindo a evolução temporal das publicações, a extração de palavras-chave via TF–IDF, a modelagem de tópicos com LDA, e a comparação vetorial com modelos de linguagem mais avançados, como o spaCy TRF. Ao final, são discutidos indicadores e métricas semânticas que reforçam a viabilidade da abordagem proposta.

4.1. Análise Temporal das Publicações

O primeiro aspecto analisado foi a distribuição temporal dos artigos coletados, com o objetivo de identificar padrões de crescimento no interesse científico pela aplicação de biomarcadores sanguíneos no diagnóstico da Doença de Alzheimer. A análise desse comportamento ao longo do tempo contribui para contextualizar a atualidade e a relevância do tema investigado. A Figura 6 apresenta a frequência absoluta de publicações por ano, considerando o período de 1988 a 2025.

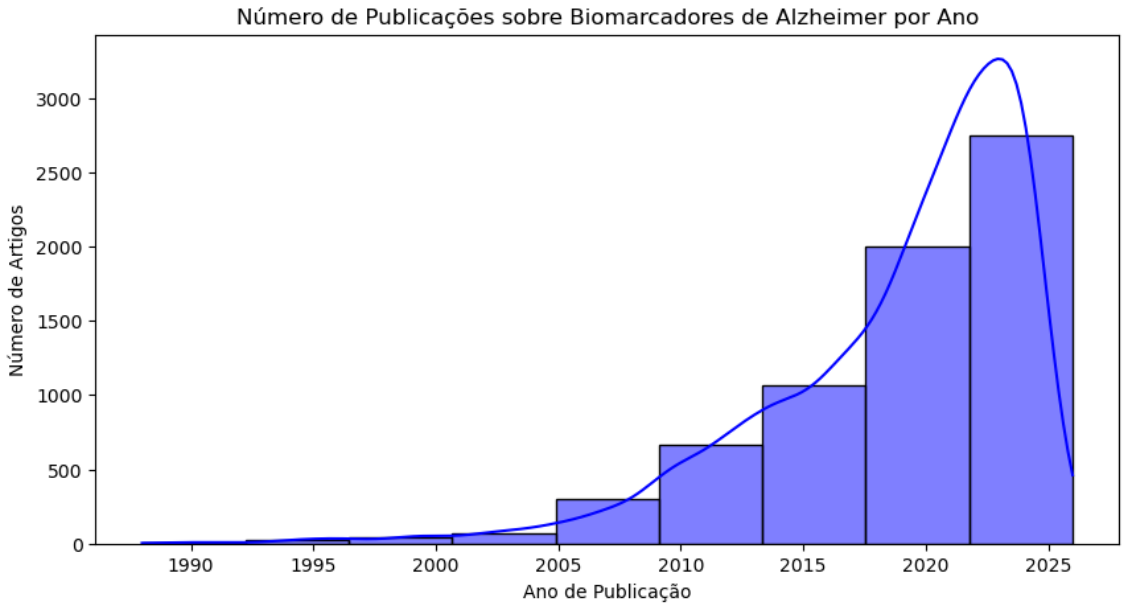

Figura 6. Histograma da frequência de publicações por ano (1988–2025).

Observa-se uma baixa quantidade de publicações até o início da década de 2010, seguida de um crescimento expressivo a partir de 2015. Esse aumento é particularmente acentuado após 2019, refletindo a ampliação do interesse na busca por diagnósticos menos invasivos e mais precoces, bem como o avanço de tecnologias aplicadas à medicina personalizada. A curva crescente sugere que o tema está em franca expansão, com tendência de manutenção no curto prazo. A Figura 7 complementa essa análise com uma visão estatística da dispersão dos dados ao longo do período.

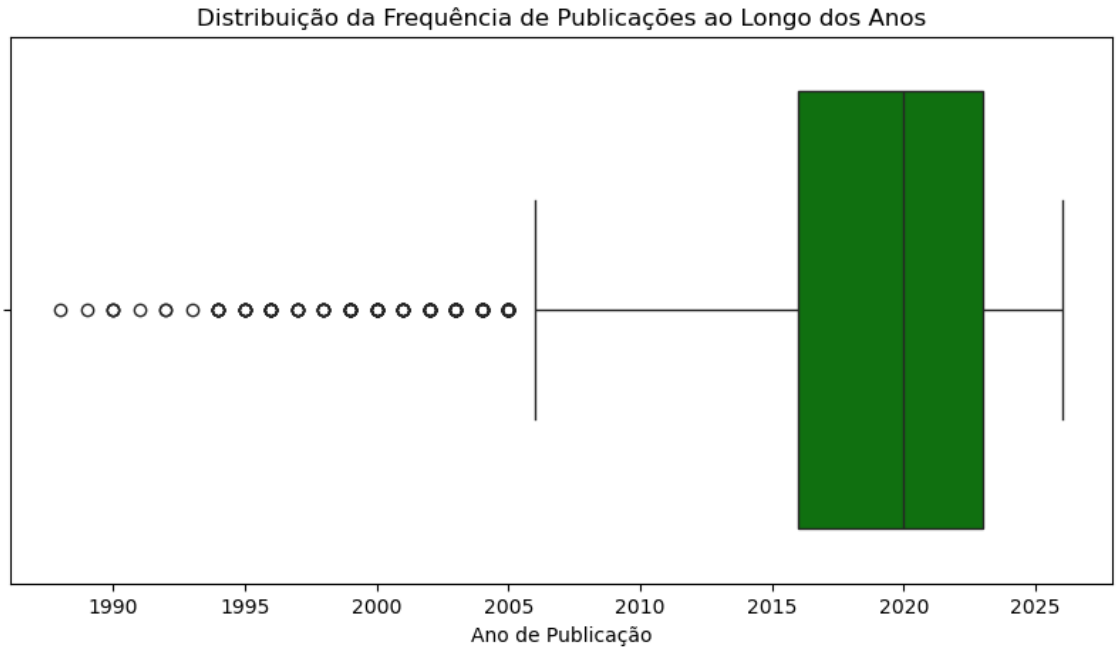

Figura 7. Boxplot da distribuição anual de publicações.

O boxplot revela que a maioria das publicações está concentrada no intervalo entre 2015 e 2023, com uma mediana por volta de 2020. Os valores mínimos e máximos indicam que, apesar de algumas publicações anteriores, foi somente na última década que o tema passou a receber atenção sistemática na literatura científica. A ausência de outliers extremos também indica que esse crescimento foi relativamente consistente, sem explosões pontuais que distorcessem a tendência.

A combinação dos dois gráficos evidencia um amadurecimento do campo nos últimos anos, justificando a aplicação de métodos automatizados para revisar um volume crescente de publicações

4.2. Aprimoramento Semântico com TF–IDF e Modelo TRF

Após a extração inicial de palavras-chave com o modelo `spaCy SM`, realizou-se uma etapa complementar com o modelo `spaCy TRF`, baseado em arquiteturas do tipo Transformer. Essa abordagem visou aprimorar a representação semântica dos textos, permitindo à vetorização TF–IDF capturar relações de contexto mais complexas e relevantes para o domínio biomédico. A Figura 8 apresenta os termos com maior peso atribuídos pelo TF–IDF após o processamento dos textos com o modelo TRF.

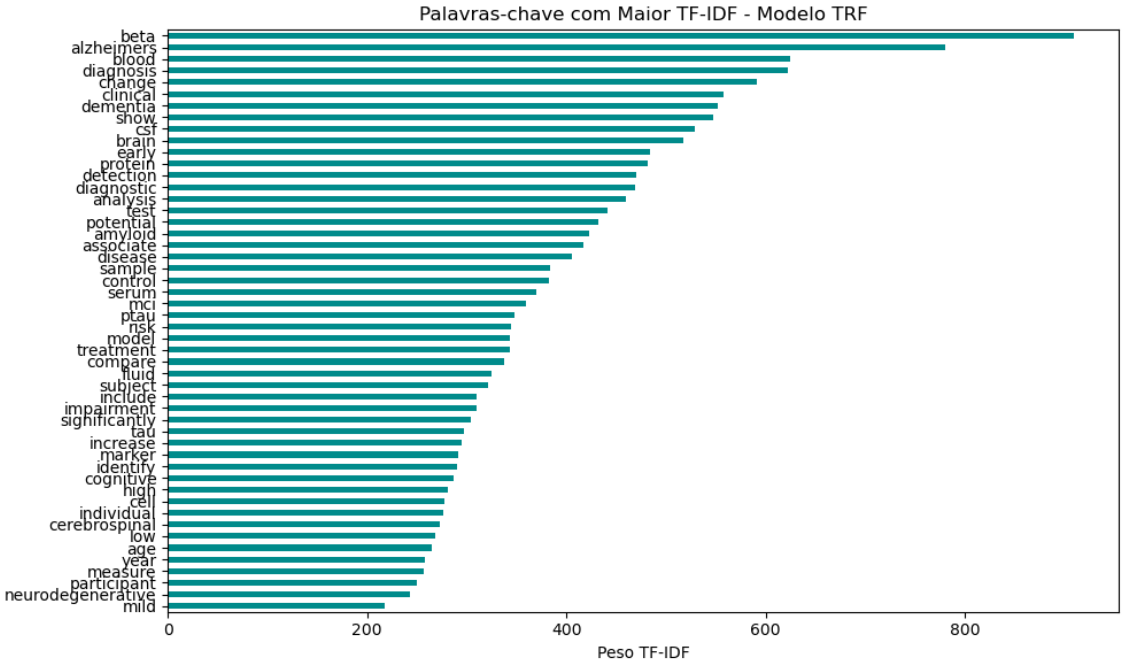

Figura 8. Palavras-chave com maior peso TF–IDF nos textos processados com o modelo `spaCy TRF`.

Observa-se que, além de manter termos já identificados anteriormente, como `diagnosis`, `cognitive`, `beta` e `biomarker`, o modelo TRF destacou com maior clareza expressões compostas semanticamente relevantes, como `cognitive impairment`, `mild cognitive`, `early diagnosis` e `disease progression`. Isso indica um ganho expressivo na capacidade do modelo de representar relações conceituais que escapam a métodos mais superficiais de pré-processamento.

Em comparação com os resultados discutidos nas figuras 4 e 5, observa-se que o modelo TRF reduziu o número de termos genéricos e aumentou a presença de bigramas com valor informativo. Esse refinamento semântico é especialmente útil na identificação de tópicos compostos e tendências específicas da literatura biomédica, como a ênfase em marcadores precoces e combinações clínicas relevantes.

Embora a execução com o modelo TRF apresente custo computacional significativamente superior (ver Tabela 5), os resultados obtidos justificam sua aplicação pontual em análises que demandam maior profundidade interpretativa.

4.3. Modelagem de Tópicos com LDA

A modelagem de tópicos foi utilizada com o objetivo de identificar agrupamentos latentes de termos nos resumos científicos analisados. Para isso, foi empregado o algoritmo Latent Dirichlet Allocation (LDA), conforme descrito na metodologia, com extração de cinco tópicos a partir da matriz TF–IDF otimizada. As Figuras 9 a 13 apresentam os termos mais representativos de cada tópico extraído, com base na distribuição de probabilidade das palavras atribuídas pelo modelo.

O agrupamento relativo ao primeiro tópico reúne termos como `samples`, `progression`, `diagnosis` e `risk`, o que sugere foco em análises clínicas relaci-

onadas à progressão da Doença de Alzheimer e à avaliação do risco com base em biomarcadores sanguíneos e liquóricos.

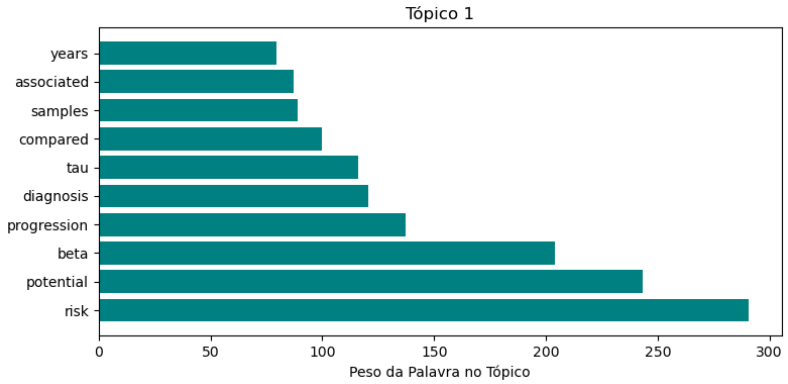

Figura 9. Tópico 1 — Biomarcadores e risco clínico

O segundo tópico apresenta forte presença de termos como `proteins`, `amyloid`, `tau` e `cerebrospinal`, remetendo a estudos tradicionais de biomarcadores como `beta-amyloid` e `p-Tau`, frequentemente analisados em fluídos como o líquor.

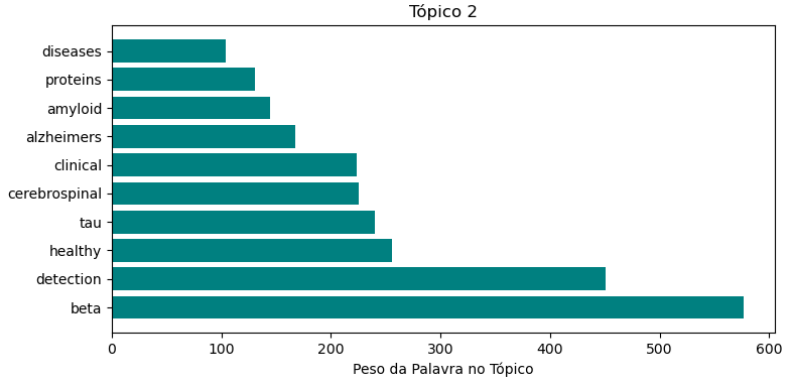

Figura 10. Tópico 2 — Proteínas, fluídos e biomarcadores clássicos

O agrupamento relativo ao tópico 3 destaca expressões como `cognitive`, `treatment`, `impairment` e `brain`, indicando foco em intervenções terapêuticas e sua relação com o declínio cognitivo e disfunções neurológicas associadas.

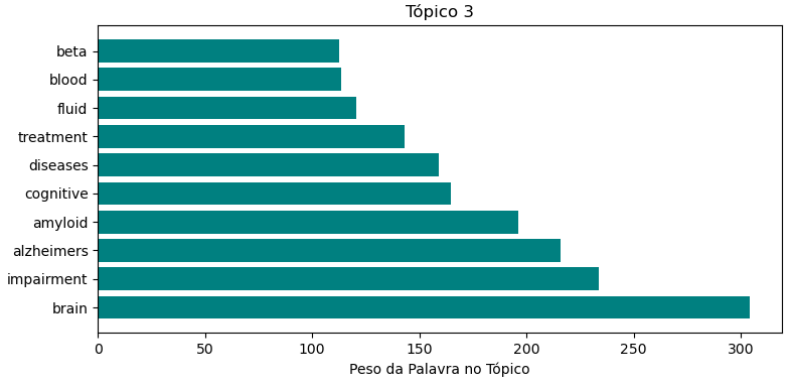

Figura 11. Tópico 3 — Tratamentos e declínio cognitivo

O tópico 4 parece reunir estudos voltados à detecção proteica em soro sanguíneo, com termos como `serum`, `protein`, `diagnostic` e `analysis`, o que sugere a presença de pesquisas laboratoriais com foco na viabilidade de exames menos invasivos.

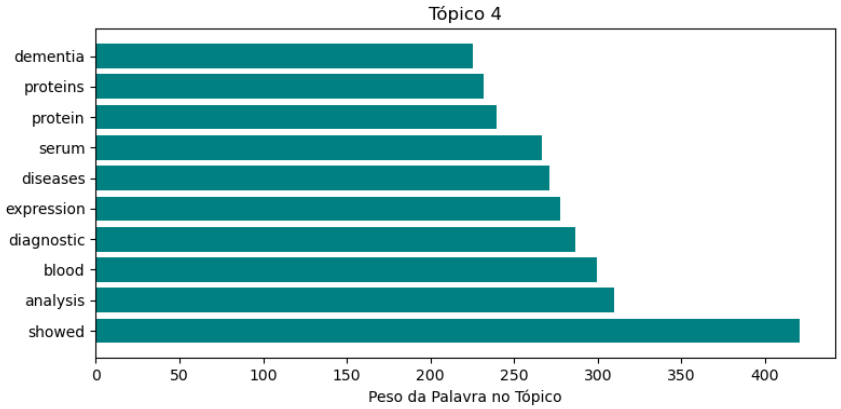

Figura 12. Tópico 4 — Sérum, expressão e análises laboratoriais

O último tópico destaca termos como `nfl`, `diagnosis`, `mild`, `controls` e `neurodegenerative`, sugerindo ênfase em estudos clínicos que avaliam pacientes com diferentes estágios de neurodegeneração, especialmente em fases iniciais da doença.

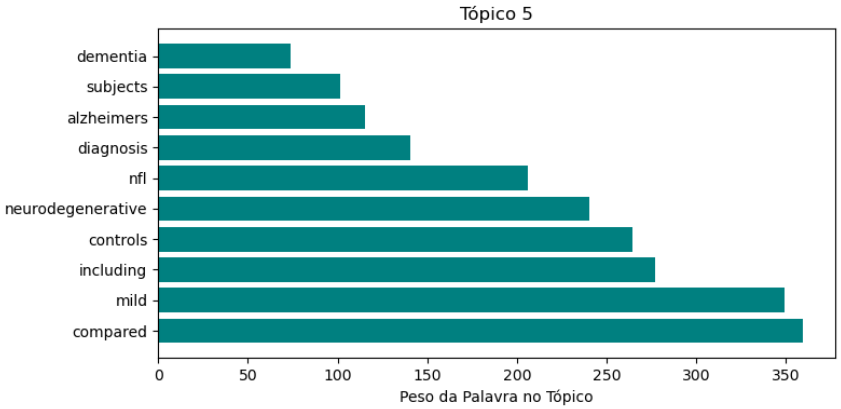

Figura 13. Tópico 5 — Neurodegeneração e diagnóstico clínico

A modelagem de tópicos via LDA demonstrou boa capacidade de segmentar os conteúdos textuais em núcleos temáticos coerentes com o domínio biomédico. Os cinco tópicos extraídos refletem áreas recorrentes na literatura recente: biomarcadores líquidos, fluídos cerebrospinais, aspectos clínicos, declínio cognitivo e expressão proteica. Esses resultados corroboram as análises anteriores por TF–IDF e reforçam o valor da modelagem de tópicos como ferramenta para revisão automatizada da literatura científica.

A Tabela 6 resume os cinco tópicos identificados, sintetizando os temas interpretados e os termos com maior recorrência em cada agrupamento. A segmentação realizada pelo modelo LDA demonstra aderência temática ao domínio biomédico, evidenciando núcleos associados a risco clínico, biomarcadores clássicos, terapias cognitivas, expressão proteica e diagnóstico precoce.

Tabela 6. Síntese dos tópicos extraídos com LDA e seus principais termos

Tópico	Tema interpretado	Principais termos
1	Biomarcadores e risco clínico	samples, progression, diagnosis, risk, potential, associated, beta, years
2	Fluídos e biomarcadores clássicos	proteins, amyloid, tau, cerebrospinal, clinical, diseases, healthy, detection
3	Declínio cognitivo e terapias	cognitive, treatment, impairment, brain, alzheimers, fluid, blood, diseases
4	Expressão e análises laboratoriais	serum, protein, diagnostic, expression, analysis, showed, subjects
5	Neurodegeneração e diagnóstico precoce	nfl, neurodegenerative, mild, controls, diagnosis, dementia, including

Esses agrupamentos complementam os achados das etapas anteriores, reforçando a consistência interna do corpus analisado. Além disso, demonstram a viabilidade da modelagem probabilística como ferramenta auxiliar para mapear tendências, identificar temas emergentes e automatizar revisões de literatura em larga escala.

4.4. Padrões Semânticos e Coocorrência de Termos

Com o objetivo de visualizar os termos mais recorrentes e semanticamente relevantes no corpus analisado, foram geradas nuvens de palavras (*wordclouds*) a partir dos textos pré-processados com os modelos `spaCy SM` e `spaCy TRF`. Cada nuvem representa os termos mais frequentes após a etapa de limpeza e normalização textual, permitindo uma avaliação qualitativa da representação semântica extraída por cada abordagem. A Figura 14 apresenta a nuvem gerada com o modelo `spaCy SM`.

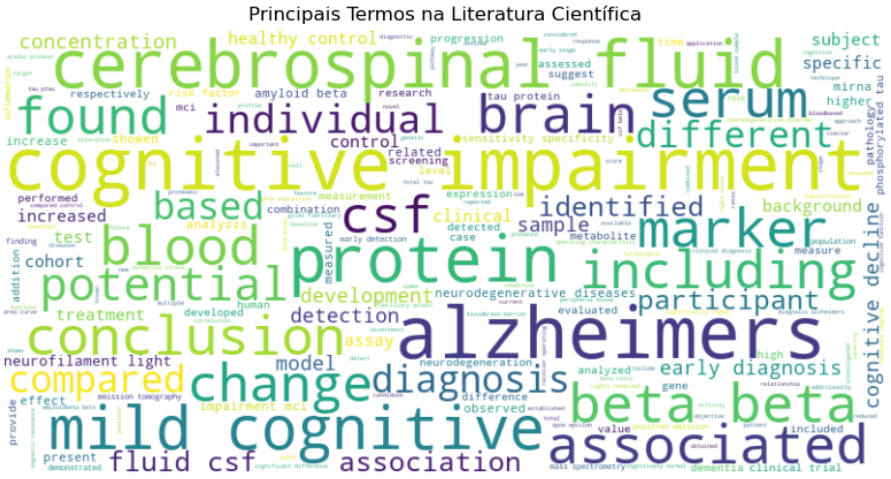

Figura 14. Nuvem de palavras com pré-processamento via `spaCy SM`.

Nessa visualização, observam-se termos centrais como `cognitive impairment`, `alzheimers`, `serum`, `protein`, `diagnosis` e `fluid`, refletindo o núcleo semântico biomédico predominante na literatura analisada. Entretanto, a

distribuição dos termos ainda mostra certa redundância e menor diversidade composicional, o que é esperado de modelos mais leves e com menor capacidade contextual. A Figura 15 mostra a nuvem gerada com o modelo `spaCy TRF`, baseado em arquiteturas do tipo Transformer.

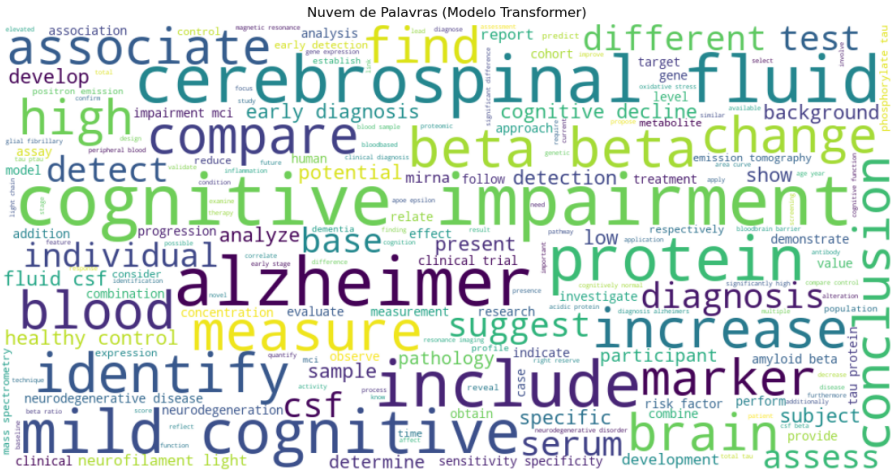

Figura 15. Nuvem de palavras com pré-processamento via `spaCy TRF`.

A nuvem baseada no modelo TRF evidencia maior variedade semântica e presença de termos compostos mais sofisticados, como `identify`, `measure`, `compare`, `include`, além da manutenção dos principais conceitos biomédicos observados anteriormente. Isso indica que a vetorização com TRF favorece a retenção de estruturas contextuais, permitindo que expressões mais complexas sejam reconhecidas com maior precisão.

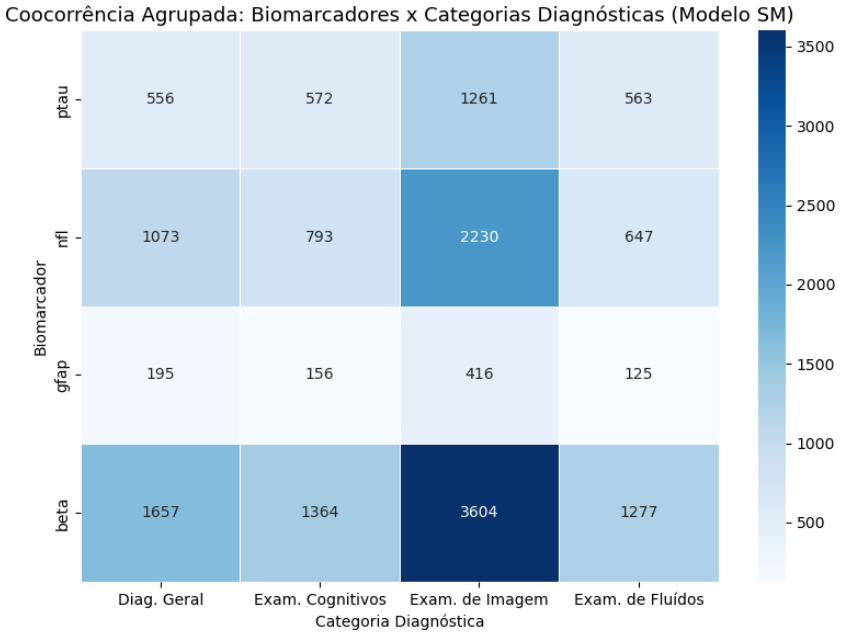

Figura 16. Coocorrência agrupada entre biomarcadores e categorias diagnósticas (Modelo SM).

A análise de coocorrência entre biomarcadores e categorias diagnósticas permitiu identificar padrões semânticos relevantes que reforçam a aderência temática do corpus ao domínio biomédico. A Figura 16 apresenta um mapa de calor construído com base no modelo `spaCy SM`, no qual os termos foram agrupados em quatro categorias diagnósticas: Diagnóstico Geral, Exames Cognitivos, Exames de Imagem e Exames de Fluídos.

Observa-se que os biomarcadores `beta` e `nfl` apresentam alta coocorrência com exames de imagem e avaliações cognitivas, o que corrobora seu papel consolidado na triagem da Doença de Alzheimer. Por outro lado, marcadores como `ptau` e `gfap` demonstram menor frequência relativa, sugerindo uma menor presença ou ênfase nos estudos incluídos neste corpus. Esses padrões de associação ajudam a mapear automaticamente como diferentes marcadores estão relacionados a abordagens diagnósticas específicas na literatura recente.

A Tabela 7 detalha os grupos de termos utilizados para compor cada categoria diagnóstica, auxiliando na interpretação semântica dos agrupamentos representados no heatmap.

Tabela 7. Categorias diagnósticas utilizadas na análise de coocorrência

Categoria	Termos utilizados	Descrição
Diagnóstico Geral	*diagnosis*, *screening*	Termos amplos relacionados ao processo de diagnóstico ou triagem de pacientes.
Exames Cognitivos	*cognitive*	Avaliações do funcionamento cognitivo, como testes de memória e raciocínio.
Exames de Imagem	*pet*, *mri*, *neuroimaging*, *fmri*, *ct*	Técnicas de imagem cerebral como PET, Ressonância Magnética e Tomografia.
Exames de Fluídos	*csf*, *blood test*, *plasma*, *elisa*	Exames laboratoriais baseados em fluídos corporais (sangue, líquor, plasma).

Essa análise de coocorrência reforça a utilidade da vetorização semântica e da categorização temática como estratégias para enriquecer a interpretação automatizada da literatura científica. Além de validar o uso de modelos como o `spaCy SM` para essa finalidade, os resultados evidenciam o potencial dessas técnicas na identificação de relações relevantes entre marcadores e métodos diagnósticos, promovendo uma compreensão estruturada e escalável do estado da arte sobre biomarcadores no diagnóstico precoce do Alzheimer.

4.5. Indicadores e Métricas de Similaridade

A análise estatística foi complementada com nove indicadores descritivos, como a evolução temporal das publicações, nuvens de palavras, frequência de biomarcadores e coocorrência com termos diagnósticos. Essas visualizações auxiliaram na compreensão das tendências da literatura científica recente.

Além disso, foram aplicadas quatro métricas quantitativas para comparar os textos processados pelos modelos `spaCy SM` e `spaCy TRF`, com o objetivo de avaliar o grau de similaridade semântica entre as versões geradas. As métricas empregadas foram:

- **Similaridade Cosseno**: mede o ângulo entre os vetores TF–IDF, refletindo a sobreposição ponderada de termos;
- **Distância de Jaccard**: avalia a interseção sobre a união dos conjuntos de tokens únicos;
- **Distância de Levenshtein**: calcula o número mínimo de edições (inserções, deleções ou substituições) necessárias para transformar um texto no outro;
- **Divergência de Kullback–Leibler (KL)**: compara distribuições de probabilidade, com suavização por `epsilon` para evitar divisão por zero.

Para avaliar a distância entre os textos e o tamanho médio dos documentos processados pelos dois modelos, a figura Figura 17 apresentar os resultados referentes a essas métricas:

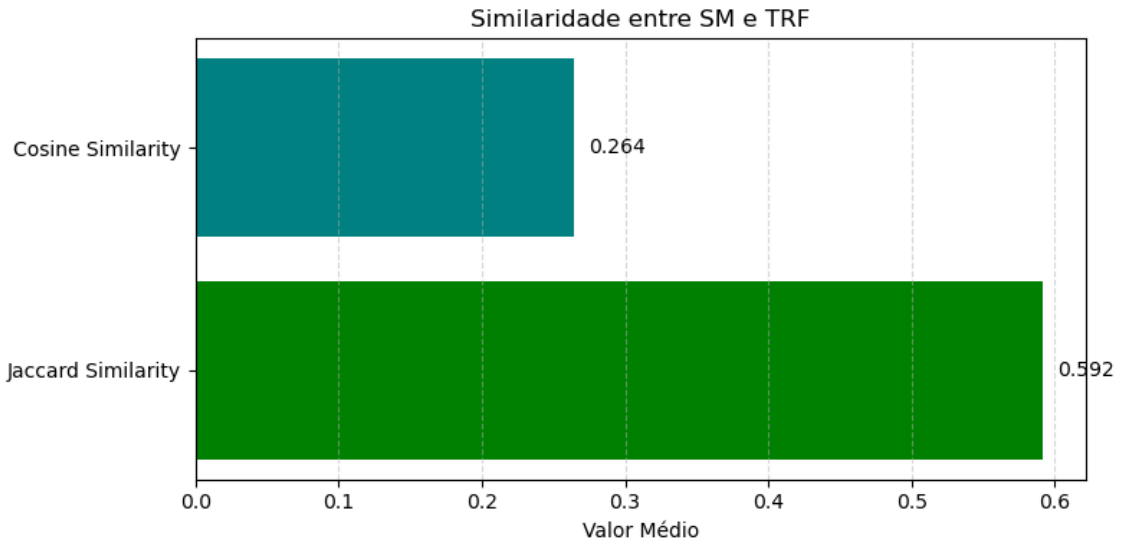

Figura 17. Distância e comprimento dos textos processados por `spaCy SM` e `spaCy TRF`.

Embora o número médio de palavras por texto seja muito semelhante entre SM e TRF (cerca de 124,5 tokens), a distância média de Levenshtein atingiu 47,4 caracteres, e a divergência KL alcançou 50,1. Isso indica que, apesar da equivalência no comprimento textual, há diferenças relevantes na estrutura lexical e na distribuição probabilística dos termos entre os dois modelos, refletindo variações na forma como cada pipeline representa semanticamente os textos.

De forma complementar, podemos observar na Figura 18 a similaridade por Cosseno e Jaccard, focando na sobreposição semântica e lexical dos textos gerados por cada abordagem. A Similaridade de Jaccard foi de 0,592, indicando uma boa interseção entre os conjuntos de tokens únicos utilizados por ambos os modelos. Já a Similaridade Cosseno foi consideravelmente menor (0,264), revelando uma baixa sobreposição na

distribuição ponderada dos termos. Esse contraste evidencia que, embora os dois modelos compartilhem parte do vocabulário, a maneira como cada um representa a importância relativa dos termos diverge substancialmente — o que justifica a utilização de ambos os pipelines em tarefas distintas de análise semântica.

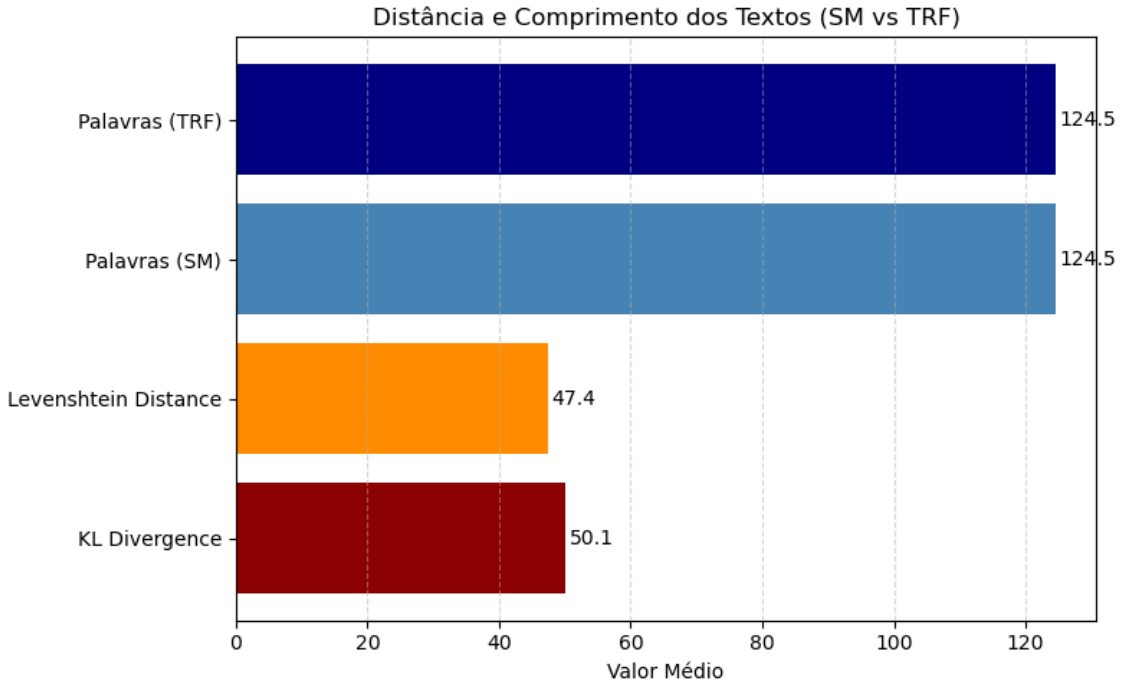

Figura 18. Similaridade entre `spaCy SM` e `spaCy TRF` com base nas métricas de Cosseno e Jaccard.

Essas métricas ajudam a quantificar diferenças que não são evidentes apenas pela leitura dos textos e reforçam a necessidade de abordagens complementares para capturar nuances semânticas mais profundas na literatura biomédica.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma abordagem automatizada para revisão sistemática da literatura científica, com foco na identificação de biomarcadores sanguíneos associados ao diagnóstico precoce da Doença de Alzheimer. Foram analisados mais de 6 mil artigos extraídos da plataforma Web of Science, utilizando técnicas de Processamento de Linguagem Natural (PLN) e modelos pré-treinados da biblioteca spaCy (`en_core_web_sm` e `en_core_web_trf`).

As técnicas de extração de palavras-chave por TF–IDF e modelagem de tópicos com LDA permitiram identificar padrões relevantes no corpus, com destaque para os biomarcadores p-Tau217, NfL, GFAP e beta-amiloide, frequentemente associados a exames como PET scan, testes cognitivos e análises de fluídos. A análise temporal revelou um crescimento expressivo das publicações a partir de 2019, evidenciando o aumento do interesse científico no tema. As métricas de similaridade e divergência aplicadas aos textos processados pelos modelos SM e TRF indicaram diferenças substanciais na representação semântica, sendo que o modelo baseado em Transformers (TRF) demonstrou maior capa-

cidade de preservação de contexto e identificação de entidades compostas, enriquecendo a análise qualitativa dos textos.

Como trabalhos futuros, aprofundar a análise quantitativa dos textos processados, investigando com mais profundidade o comportamento de métricas de comparação como a Similaridade Cosseno, que mede o ângulo entre os vetores TF–IDF e reflete a sobreposição ponderada de termos; a Distância de Jaccard, que avalia a interseção relativa entre os conjuntos de tokens únicos; a Distância de Levenshtein, baseada no número mínimo de edições necessárias para transformar um texto no outro; e a Divergência de Kullback–Leibler (KL), que compara distribuições de probabilidade suavizadas para evitar distorções. Tais métricas oferecem diferentes perspectivas sobre a semelhança lexical e semântica entre versões textuais, sendo promissoras para a construção de indicadores sintéticos mais robustos.

Além disso, sugere-se a ampliação da análise para outras bases de dados científicas, como PubMed e Scopus, bem como a aplicação de algoritmos supervisionados para classificação automática dos artigos por temática ou tipo de biomarcador. Também está prevista a adaptação do pipeline para o idioma português, a incorporação de embeddings contextuais como BERT ou BioBERT, e o refinamento dos componentes de reconhecimento de entidades nomeadas (NER) com foco em terminologia biomédica. Os resultados obtidos demonstram a viabilidade do uso de PLN como ferramenta de apoio à revisão científica em larga escala, contribuindo para o mapeamento de tendências emergentes e subsidiando decisões em contextos clínicos e acadêmicos.

Agradecimentos

Os autores agradecem à Samsung Eletrônica da Amazônia Ltda., por meio do Projeto Aranouá, e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro por meio do Programa de Excelência Acadêmica (PROEX). Este trabalho é resultado do projeto de Pesquisa e Desenvolvimento (P&D) 001/2021, firmado com o Instituto Federal do Amazonas e a FAEPI, com financiamento da Samsung.

Referências

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003a). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chen, D., Wang, J., and Zhang, F. (2023a). Blood-based biomarkers for alzheimer's disease: current state and future directions. *Nature Reviews Neurology*, 19(6):357–377.

Chen, D., Xiao, D., Blevins, D., and Bateman, R. J. (2023b). Blood-based biomarkers for alzheimer's disease: current state and future directions. *Nature Reviews Neurology*.

Cui, L. and Lee, D. (2020). Natural language processing and biomedical literature mining in translational bioinformatics. *Yearbook of Medical Informatics*, 29(01):138–145.

Cui, L. e. a. (2020). Natural language processing and its applications in biomedical domain. *Briefings in Bioinformatics*, 21(1):160–179.

González-Castro, V., Palacios, R., and López, M. (2021). Text mining in biomedical literature: a systematic review. *Int. J. Environ. Res. Public Health*, 18(14):7419.

Hampel, H., O'Bryant, S. E., Molinuevo, J. L., and et al. (2018). Blood-based biomarkers for alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, 14:639–652.

Honnibal, M. and Montani, I. (2023). spacy: Industrial-strength natural language processing in python. `https://spacy.io/models`. Acesso em: abr. 2025.

Honnibal, M., Montani, I., Landeghem, S. V., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python. *Zenodo*.

Karikari, T. K. e. a. (2020). Blood phosphorylated tau 181 as a biomarker for alzheimer's disease: a diagnostic performance and prediction modelling study using data from four prospective cohorts. *The Lancet Neurology*, 19(5):422–433.

Palmqvist, S. e. a. (2020). Discriminative accuracy of plasma phospho-tau217 for alzheimer disease vs other neurodegenerative disorders. *JAMA*, 324(8):772–781.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2023). Scikit-learn: Machine learning in python - latentdirichletallocation. `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html`. Acesso em: abr. 2025.

Preische, O. e. a. (2019). Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic alzheimer's disease. *Nature Medicine*, 25(2):277–283.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*. Rutgers University.

Thijssen, E. H. e. a. (2020). Diagnostic value of plasma phosphorylated tau181 in alzheimer's disease and frontotemporal lobar degeneration. *Nature Medicine*, 26(3):387–397.

World Health Organization (2022). Dementia. *Fact Sheets*.