PETROVEC: DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS VETORIAIS DE PALAVRAS EM PORTUGUÊS PARA O DOMÍNIO DE ÓLEO E GÁS

Diogo da Silva Magalhães Gomes

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Alexandre Gonçalves Evsukoff

Rio de Janeiro

Março de 2021

PETROVEC: DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS VETORIAIS DE PALAVRAS EM PORTUGUÊS PARA O DOMÍNIO DE ÓLEO E GÁS

Diogo da Silva Magalhães Gomes

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Orientador: Alexandre Gonçalves Evsukoff

Aprovada por: Prof. Alexandre Gonçalves Evsukoff
Prof. Nelson Francisco Favilla Ebecken
Prof. Leonardo Alfredo Forero Mendoza
Profª. Viviane Pereira Moreira
Dr. Rafael Jesus de Moraes

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 2021

Gomes, Diogo da Silva Magalhães

PetroVec: desenvolvimento e avaliação de modelos vetoriais de palavras em português para o domínio de óleo e gás / Diogo da Silva Magalhães Gomes. – Rio de Janeiro: UFRJ/COPPE, 2021.

XIV, 123 p.: il.; 29,7 cm.

Orientador: Alexandre Gonçalves Evsukoff

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2021.

Referências Bibliográficas: p. 107-120.

1. PLN. 2. embeddings. 3. vetorização. I. Gonçalvez, Alexandre Evsukoff. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Para meus pais Leonor e Luis, meus verdadeiros mestres.

Aos meus filhos Larissa e Pedro, minha força motivadora e fontes de inspiração que iluminam meus dias.

À minha amada esposa Roberta, pela incrível parceria nessa jornada da vida.

Agradecimentos

Empreender o projeto de um curso de doutoramento é ingressar em uma longa e desafiadora jornada, cujo sucesso não teria sido possível sem o suporte, cooperação e parceria de inúmeras pessoas que me inspiraram a superar os desafios nessa caminhada.

Primeiramente, expresso minha gratidão pela inestimável colaboração do meu orientador, prof. Alexandre Evsukoff. Seu aconselhamento motivador, perspicácia, objetividade e companheirismo foram essenciais para alcançar o cumprimento dos objetivos de pesquisa. Agradeço também a todo o Programa de Engenharia Civil da COPPE/UFRJ, pela oportunidade e pelo privilégio de fazer parte deste excelente grupo.

Agradeço a todos que participaram na preparação dos objetivos de pesquisa desta tese, compartilhando uma valiosa rede de colaboração motivada a desbravar esse domínio técnico ainda pouco explorado sob a temática da linguística computacional. Em especial, agradeço imensamente pela valiosa dedicação dos alunos, pesquisadores e professores das Universidades UFRGS, PUC-RS e PUC-Rio.

Aos colegas da TIC e do Centro de Pesquisas e Desenvolvimento da Petrobras (Cenpes), que colaboraram imensamente para tornar este trabalho possível. Em especial, agradeço aos queridos amigos do projeto Busca Semântica, pela agradável convivência diária e pelas inspiradoras discussões técnicas que nos desafiam a uma busca contínua pelo conhecimento científico: Regis Kruel, Max, Eugênio, Marcelo Rezende, Rafael, Luciana e Vitor Batista. Agradeço especialmente aos amigos doutorandos, Fábio Cordeiro e Antônio Marcelo, pela virtuosa parceria, amizade e cumplicidade no cumprimento das nossas jornadas acadêmicas.

Meus agradecimentos à liderança da Petrobras por incentivar e permitir a minha participação nesta especialização acadêmica. À gerente Lourdes Alice pelo perseverante e valoroso incentivo para viabilizar a minha inscrição no doutorado. Ao gerente Miccolis, pelas palavras motivadoras e pelo companheirismo ao oferecer as condições favoráveis para a condução da minha pesquisa. Ao gerente Luiz Carlos por permitir e incentivar a continuidade e conclusão deste trabalho. Em especial, meu sincero agradecimento ao amigo Flávio Gondim, por seus inúmeros conselhos e por seu admirável exemplo de

liderança, empatia e encorajamento, que foram fundamentais para a condução desta pesquisa.

Por fim, guardo especial e profunda gratidão à minha família, pelo apoio incondicional e motivação para que eu pudesse me dedicar aos estudos. À minha esposa Roberta, minha melhor amiga desde sempre, que com muita paciência, amor e extraordinária alegria me incentivou a superar os desafios. Aos meus filhos Larissa e Pedro, pela inocente compreensão perante os momentos mais adversos. Seus sorrisos e gargalhadas são a recompensa diária e principal força motivadora para seguir adiante.

Aos meus pais Leonor e Luis, por seu amor imenso e pela abnegada dedicação para propiciar as melhores condições para a nossa educação. São minhas maiores inspirações, meus referenciais de sabedoria, caráter, honestidade e respeito. Aos meus queridos irmãos Danilo e Amanda, por seus admiráveis exemplos de personalidade, inteligência, companheirismo e amizade. À minha avó Alexandrina, matriarca dessa extraordinária família, exemplo de dedicação, sabedoria e amor imensurável. À minha madrinha Tia Isa (*in memorian*), pelo eterno amor e carinho, sempre presentes desde a minha infância.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

PETROVEC: DESENVOLVIMENTO E AVALIAÇÃO DE MODELOS VETORIAIS DE PALAVRAS EM PORTUGUÊS PARA O DOMÍNIO DE ÓLEO E GÁS

Diogo da Silva Magalhães Gomes

Março/2021

Orientador: Alexandre Gonçalves Evsukoff
Programa: Engenharia Civil

Este trabalho apresenta o **PetroVec**: um conjunto de modelos vetoriais de palavras em português especializados no domínio de Óleo e Gás. Para viabilizar o treinamento dos modelos, criamos um corpus representativo do domínio, composto por uma extensa coleção de documentos técnicos e acadêmicos publicados em português por Universidades e instituições de referência na indústria nacional de petróleo. O corpus especializado contempla mais de 85 milhões de tokens e representa o maior conjunto textual público atualmente reportado na literatura científica para o domínio de Óleo e Gás. Os modelos são submetidos a uma abrangente cobertura de avaliações, contemplando metodologias quantitativas baseadas em análises intrínseca e extrínseca, além de uma série de análises qualitativas para explorar propriedades linguísticas codificadas no espaço semântico dos modelos. A análise intrínseca foi realizada a partir da criação de um *dataset* de similaridade semântica composto por 1500 pares de termos anotados por especialistas em geociências, enquanto a análise extrínseca consistiu na aplicação prática dos modelos em uma tarefa de reconhecimento de entidades nomeadas no subdomínio de geologia. Adicionalmente, realizamos análises comparativas dos nossos resultados em relação a um modelo público de contexto geral de referência em português. Nossas análises convergem ao evidenciar que os modelos **PetroVec** apresentam resultados consistentemente superiores ao modelo público de referência em todas as avaliações, sugerindo que os modelos especializados são capazes de automaticamente capturar propriedades sintáticas e semânticas específicas do vocabulário técnico de domínio de maneira não-supervisionada a partir do corpus de treinamento.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

PETROVEC: DEVELOPMENT AND EVALUATION OF PORTUGUESE WORD EMBEDDING MODELS FOR THE OIL AND GAS DOMAIN

Diogo da Silva Magalhães Gomes

March / 2021

Advisor: Alexandre Gonçalves Evsukoff
Department: Civil Engineering

This work presents **PetroVec**, a set of word embedding models in Portuguese for the O&G domain. To make model training feasible, we created a specialized corpus composed of a vast collection of technical and scientific documents, published in Portuguese by Universities and major institutions from national petroleum-related industry. The specialized corpus comprises about 85 million tokens and it is currently the largest public textual resource ever reported for the O&G domain. Our specialized models are thoroughly evaluated, comprising quantitative methodologies based on intrinsic and extrinsic approaches, in addition to a series of qualitative analyses to explore linguistic properties encoded in the models' semantic space. The intrinsic evaluation is performed by creating a semantic similarity dataset composed of 1500 pairs of terms labeled by experts in geosciences, whereas the extrinsic evaluation consists of a downstream task for named entity recognition in the Geology subdomain. Furthermore, we conducted a comprehensive analysis comparing our models and a pre-trained general-domain model in Portuguese. Our findings confirm that **PetroVec** consistently outperforms the general-context reference model, suggesting that our models were able to automatically capture syntactic and semantic vocabulary-specific properties from the specialized training corpus.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de siglas e abreviaturas

API	Interface de Programação de Aplicativos (*Application Programming Interface*)
ANP	Agência Nacional de Petróleo, Gás e Biocombustíveis
BERT	*Bidirectional Encoder Representations from Transformers*
CENPES	Centro de Pesquisas e Desenvolvimento da Petrobras
CNN	Redes Neurais Convolucionais *(Convolutional Neural Networks)*
COPPE	Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
CRF	*Conditional Random Fields*
E&P	Exploração e Produção
HMM	*Hidden Markov Model*
HTTP	*Hypertext Transfer Protocol*
LSTM	*Long Short-term memory*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
O&G	Óleo e Gás
PLN	Processamento de Linguagem Natural
POS	*Part-of-speech*
PUC-RS	Pontifícia Universidade Católica do Rio Grande do Sul
RecNN	Redes Neurais Recursivas *(Recursive Neural Networks)*
REN	Reconhecimento de Entidades Nomeadas
RNA	Redes Neurais Artificiais
RNN	Redes Neurais Recorrentes *(Recurrent Neural Networks)*
SOC	Sistema de Organização do Conhecimento *(Knowledge Organization System)*
SOG	*Schlumberger Oilfield Glossary*
SPE	*Society of Petroleum Engineers*
SVM	*Support Vector Machine*
UFRJ	Universidade Federal do Rio de Janeiro
UFRGS	Universidade Federal do Rio Grande do Sul
WE	*Word Embeddings*

Capítulo I

Introdução

"*We make our world significant by the courage of our questions and the depth of our answers.*"

Carl Sagan

Este capítulo apresenta uma visão geral da tese, oferecendo uma contextualização sobre o tema de pesquisa, as motivações, caracterização do problema e objetivos da pesquisa, além das principais contribuições e publicações realizadas pelo autor no decorrer do curso de doutoramento. Por fim, apresenta-se a estrutura geral de organização do texto que será utilizada no decorrer da tese.

1.1 Introdução

Na era da chamada Transformação Digital (HENRIETTE et al., 2017), essencialmente caracterizada pela rica disponibilidade de técnicas de ciência de dados e inteligência artificial, a indústria de Óleo e Gás (O&G) tem sido continuamente desafiada a fazer melhor uso das informações atualmente disponíveis em suas imensas bases de dados (WORLD ECONOMIC FORUM, 2017). Muitas companhias vêm reforçando seus investimentos em tecnologias digitais no intuito de obter um melhor aproveitamento desses dados e, assim, melhorar a eficiência em suas operações e processos decisórios (MATT, HESS e BENLIAN, 2015).

No decorrer das últimas décadas, motivados pela diminuição de custos de *hardware* de armazenamento e processamento, além do surgimento de técnicas e plataformas que viabilizaram o processamento eficiente de grandes volumes de dados (Big Data), um novo paradigma se estabeleceu no sentido maximizar a diversidade de informações coletadas e armazenadas, caracterizando um crescimento exponencial do volume de dados capturados. Estima-se que 80% dos dados armazenados em bases corporativas corresponda a formatos não-estruturados (BLINSTON e BLONDELLE, 2017), como texto, imagens e áudio.

Dispersas nesses imensos volumes de dados, há importantes informações que, se corretamente identificadas e processadas, podem fornecer relevantes insumos para inúmeros processos decisórios, além de dar suporte a uma ampla variedade de atividades acadêmicas e industriais (ITTOO *et al.*, 2016). Referindo-se especificamente a dados no formato texto, há um enorme manancial de informações valiosas e não plenamente utilizadas, imersas em relatórios técnicos, artigos científicos, logs de operação, pareceres técnicos, análises laboratoriais, periódicos, entre outros documentos.

Embora essa rica disponibilidade de dados represente uma oportunidade para as corporações, obter informações relevantes extraídas a partir desses imensos volumes torna-se também um desafio. Muitas instituições não fazem uso do pleno potencial dos dados que possuem e, portanto, informações estratégicas muitas vezes tornam-se negligenciadas por não serem adequadamente identificadas e processadas (FORBES, 2017). Face à intensa competitividade no cenário industrial, é economicamente vital que as empresas de O&G sejam capazes de plenamente identificar e utilizar as informações

disponíveis em suas bases de dados, a fim de maximizar sua eficiência operacional (BLINSTON e BLONDELLE, 2017; LU *et al.*, 2019).

Nesse contexto, recentes avanços nas áreas de Processamento de Linguagem Natural (PLN) e aprendizagem profunda (*deep learning*) (LECUN *et al.*, 2015; GOODFELLOW *et al.*, 2016) têm contribuindo significativamente para se obter um melhor aproveitamento do potencial de uso dessas informações, com o surgimento de novas técnicas para viabilizar o processamento e extração de conhecimento a partir de bases de dados textuais (YOUNG *et al.*, 2018; KHURANA *et al.*, 2018). Muitos desses métodos têm sido utilizados com sucesso em diversas aplicações industriais (ITTOO *et al.*, 2016; BLINSTON e BLONDELLE, 2017; NOORALAHZADEH *et al.*, 2018; CORDEIRO *et al.*, 2019).

Algoritmos de PLN utilizam texto não-estruturado como dados de entrada e, portanto, é crucial obter representações matemáticas adequadas e significativas para seus elementos textuais. Modelos de vetorização de palavras, ou *word embeddings (WE)*[1], são um dos principais fundamentos utilizados por aplicações de PLN, e consistem em aplicar métodos não-supervisionados a partir de conjuntos de dados textuais, a fim de atribuir uma representação vetorial densa *n*-dimensional para cada termo do vocabulário (BENGIO *et al.*, 2013). Esses modelos fundamentam-se na hipótese distribucional (SAHLGREN, 2008), caracterizada por assumir que palavras que aparecem em um mesmo contexto tendem a possuir significados semelhantes. Portanto, os WE são capazes de capturar características essenciais de linguagem, como morfologia, sintaxe e, em especial, semântica (MIKOLOV *et al.*, 2013a, 2013b; CAMACHO-COLLADOS e PILEHVAR, 2018). Esses modelos de representação vetorial são uma das unidades fundamentais em qualquer aplicação de PLN, contribuindo para otimizar o desempenho dessas aplicações devido à sua grande capacidade de generalização (GOLDBERG, 2016), e são utilizados em diversos casos práticos como tradução automática, classificação de texto, sistemas de perguntas e respostas, entre outros (CAMACHO-COLLADOS e PILEHVAR, 2018).

O desenvolvimento de modelos vetoriais especializados demanda a utilização de grandes conjuntos de dados em formato textual, denominados corpora, adequadamente

[1] Neste trabalho, os termos 'modelos vetoriais de palavras', 'vetorização de palavras', '*word embeddings* (WE)' ou simplesmente 'modelos, serão igualmente utilizados para designar o mesmo conceito, referindo-se ao objeto de estudo principal representado pelos modelos PetroVec

preparados para o treinamento dos algoritmos de aprendizagem automática, o que nem sempre está disponível em quantidade e qualidade suficientes para um determinado contexto. Nesses cenários, técnicas de transferência de aprendizado (*transfer learning*) (RUDER, 2019) são comumente utilizadas, e consistem em obter modelos originalmente pré-treinados a partir de conjuntos de dados de contexto geral, e reutilizá-los para alimentar aplicações em uma tarefa específica (CER *et al.*, 2018). Há uma ampla variedade de modelos de WE pré-treinados em corpora de contexto geral (FARES *et al.*, 2017), para diversos idiomas. Para o português, alguns poucos trabalhos se propuseram a conduzir estudos para geração de modelos no idioma (RODRIGUES e BRANCO, 2016; HARTMAN *et al.*, 2017; RODRIGUES e BRANCO, 2018). Entretanto, estudos recentes apresentam consistentes evidências de que o desenvolvimento de modelos vetoriais especializados, *i.e.* treinados a partir de corpora específicos do domínio e idioma em que serão aplicados, pode melhorar significativamente a qualidade de representação semântica dos termos e, consequentemente, o desempenho dos algoritmos de PLN em que serão utilizados (LAI *et al.*, 2016; DIAZ *et al.*, 2016; PAKHOMOV *et al.*, 2016; NOORALAHZADEH *et al.*, 2018; WANG *et al.*, 2018, ALSENTZER *et al.*, 2019, TSHITOYAN *et al.*, 2019; PADARIAN e FUENTES, 2019; GOMES *et al.*, 2021).

Nesse sentido, o domínio de O&G possui reconhecidas particularidades semânticas para representação do seu vocabulário técnico altamente especializado, conforme detalhado na Seção 1.2. Isto é, no contexto desse domínio específico, há inúmeros termos que assumem significados muito peculiares que os diferenciam de textos comuns, motivando, portanto, o desenvolvimento de modelos de *embeddings* especializados. Como um dos principais demandantes dessas soluções tecnológicas, o pré-sal brasileiro representa uma importante fronteira exploratória para a indústria de O&G, cujo potencial tem atraído grande interesse internacional para investimentos em projetos de Exploração e Produção (CLAVIJO *et al.*, 2019). Além disso, a maior parte de sua documentação está em português. Entretanto, apesar do português ser um dos idiomas com o maior número de falantes nativos e da sua importância para a indústria de O&G mundial, a disponibilidade de corpora e de modelos especializados para o domínio nesse idioma é muito escassa na literatura científica.

Portanto, a fim de preencher essa lacuna, este trabalho apresenta o **PetroVec**: um conjunto de modelos de vetorização de palavras em português especializados no domínio de Óleo e Gás. Os modelos são treinados em um grande corpus representativo do domínio, criado a partir da coleta de milhares de documentos técnicos e acadêmicos publicados em português pelas principais instituições de referência na área de petróleo, contemplando artigos científicos, teses, dissertações, relatórios técnicos e periódicos. O corpus especializado é composto por cerca de 85,7 milhões de tokens, e representa o maior conjunto textual reportado para o domínio de O&G neste idioma. Os modelos vetoriais são treinados utilizando os algoritmos Word2Vec (MIKOLOV *et al.*, 2013a) e FastText (BOJANOWSKI *et al.*, 2017), em duas variações explorando diferentes composições dos corpora especializados.

Para verificar a qualidade dos modelos **PetroVec**, realizamos uma abrangente cobertura de testes e avaliações, contemplando metodologias quantitativas com análises intrínseca e extrínseca, além de uma série de análises qualitativas. Face à escassez de dados anotados disponíveis na literatura científica, conduzimos uma iniciativa em parceria com o Centro de Pesquisa e Desenvolvimento da Petrobras (Cenpes) e com pesquisadores das Universidades UFRGS e PUC-RS para viabilizar a realização das avaliações intrínseca e extrínseca. Para a *avaliação intrínseca*, criamos um *dataset* contendo índices de similaridade semântica para 1500 pares de termos técnicos relacionados ao domínio, anotados por especialistas em geociências. Para a *avaliação extrínseca*, realizamos uma metodologia para avaliar a qualidade dos modelos quando aplicados a uma tarefa prática de Reconhecimento de Entidades Nomeadas (REN) no domínio de geociências.

Desta forma, os modelos são detalhadamente avaliados, contemplando metodologias quantitativas baseadas em análise intrínseca e extrínseca, além de abordagens qualitativas, a partir de analogias de palavras, coerência de espaço semântico e categorização de conceitos. Adicionalmente, os resultados são comparativamente analisados em relação a um modelo público de domínio genérico em português, fornecido por HARTMANN *et al.* (2017), além de um modelo especializado desenvolvido em um trabalho preliminar do autor desta tese (GOMES *et al.*, 2018), para servirem como base referência para as métricas. As avaliações confirmam que os modelos especializados **PetroVec** superam consistentemente o modelo genérico e o modelo *baseline*,

evidenciando sua capacidade de capturar nuances semânticas de representação do vocabulário técnico de O&G a partir do corpus de treinamento.

Todos os artefatos desenvolvidos nesta tese estão disponíveis para a comunidade científica no repositório público **Petrolês**[2] (Figura 1.1). O ambiente Petrolês, desenvolvido pelo Centro de Pesquisas da Petrobras em parceria com a PUC-Rio, é uma iniciativa nacional pioneira com o objetivo de compartilhar recursos de PLN especializados no domínio de O&G para o idioma português, buscando servir como referência para promover o desenvolvimento de novas soluções junto à indústria e grupos de pesquisa em inteligência artificial atuantes nesse setor. No Petrolês estão disponíveis os corpora e os modelos vetoriais **`PetroVec`**, e no repositório GitHub[3] estão publicados os *datasets* anotados de validação e todo o código-fonte desenvolvido para o pré-processamento, treinamento e avaliação dos modelos. Espera-se que tanto a comunidade científica como a indústria de Óleo e Gás possam se beneficiar com a disponibilidade dos produtos desenvolvidos nesta tese, possivelmente contribuindo para fomentar novas pesquisas em PLN aplicadas na área de petróleo, cuja literatura encontra-se carente de tais insumos.

Figura 1.1: Petrolês: repositório público de artefatos de PLN para o setor de O&G, onde estão disponibilizados os modelos PetroVec e os corpora de domínio

[2] Petrolês - Repositório público de artefatos de PLN para a indústria de O&G em português. Disponível em: http://petroles.ica.ele.puc-rio.br/
[3] Repostório Github do PetroVec: https://github.com/Petroles/Petrovec

Por fim, é relevante mencionar que esta pesquisa integra uma ampla iniciativa de colaboração interinstitucional, conduzida pelo Centro de Pesquisas da Petrobras em parceria com algumas das principais Universidades brasileiras, como COPPE/UFRJ, UFRGS, PUC-RS e PUC-Rio, envolvendo um grupo de pesquisadores dedicados a desenvolver novas tecnologias para avançar o estado-da-arte em soluções de PLN para o domínio de O&G em português. O diagrama apresentado na Figura 1.2 ilustra as principais etapas realizadas neste trabalho, contemplando os processos de criação de corpora, pré-processamento dos dados, treinamento e avaliação dos modelos **`PetroVec`**, assim como os principais parceiros institucionais.

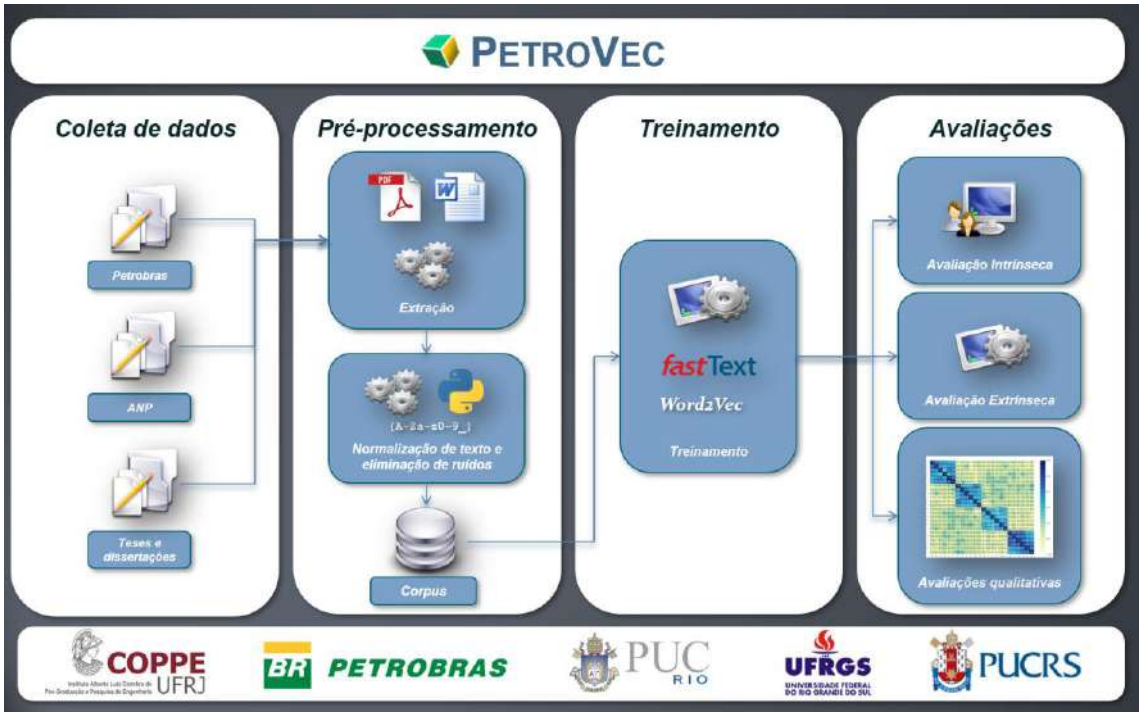


Figura 1.2: Diagrama das principais etapas envolvidas no desenvolvimento dos modelos `PetroVec` e do portal Petrolês, listando os principais parceiros institucionais

1.2 Caracterização do problema

O vocabulário estritamente técnico da área de Óleo e Gás (O&G) representa um desafio para algoritmos de PLN. Alguns termos podem assumir significados muito distintos em relação ao senso comum (GOMES et al, 2018, 2021; NOORALAHZADEH *et al.*, 2018; CORDEIRO *et al.*, 2019), como *'arvore de natal'* (um conjunto de válvulas conectadas ao topo do poço), *'falha'* (uma quebra ou fratura na superfície rochosa, em que se observa um deslocamento relativo dos blocos formados), ou nomes técnicos e

siglas de equipamentos como '*BOP'* (acrônimo para *blowout preventer*, preventor de erupção, equipamento de segurança capaz de vedar a passagem de óleo ou gás em situações emergenciais).

A Tabela 1.1 detalha um conjunto de exemplos de casos típicos em que termos do vocabulário técnico de O&G assumem um significado completamente distinto de seu senso comum, em comparação a um contexto genérico. Para compor essas definições, reunimos um conjunto de termos técnicos e utilizamos duas fontes de referência para caracterizar o seu significado: o dicionário Michaelis[4] do idioma português, para as definições de contexto geral; e o Dicionário do Petróleo em Língua Portuguesa (FERNÁNDEZ *et al.*, 2009)[5] para as definições no domínio técnico. Ressalta-se que, para todos os termos, em especial no domínio genérico, é comum haver mais de uma definição (polissemia). Portanto, nesses casos, somente a primeira definição foi considerada. Cabe destacar, ainda, os inúmeros casos em que os termos técnicos são específicos apenas para a área de O&G, não existindo, portanto, em vocabulários comuns. Esses termos não possuem representação equivalente em modelos vetoriais de contexto geral, e representam um problema denotado como palavras fora-do-vocabulário (*out-of-vocabulary*, OOV), apesar de serem de fundamental importância para o bom desempenho de algoritmos de PLN que precisem considerá-los em suas análises, o que reforça a motivação pelo desenvolvimento de modelos especializados. A Tabela 1.2 reúne exemplos utilizados com frequência em documentos técnicos, selecionados a partir do Siglário[6] publicado pelo Dicionário do Petróleo, para os quais não existe correspondência no Dicionário Michaelis.

Portanto, a especificidade técnica do domínio de O&G justifica a demanda pelo desenvolvimento de modelos especializados, capazes de capturar as nuances de propriedades semânticas e sintáticas de forma a manter a correta capacidade de generalização dos modelos de PLN. Entretanto, na literatura científica, tais modelos especializados não estão disponíveis para o domínio de O&G em Português.

[4] https://michaelis.uol.com.br/moderno-portugues/
[5] O Dicionário do Petróleo disponibiliza uma versão online gratuita: http://dicionariodopetroleo.com.br/. Acesso em 01/05/2020
[6] http://dicionariodopetroleo.com.br/siglario/

Tabela 1.1: Definições de termos para um contexto genérico e para o domínio técnico de O&G

Termo	Definição – Dicionário Michaelis	Definição – Dicionário do Petróleo
Braço	*Cada um dos membros superiores do corpo humano.*	*Mola em arco ou alavanca articulada a uma sonda de perfilagem para pressioná-la contra a parede do poço e desse modo manter o patim ou almofada;*
Calado	*Que não diz nada ou que fica em silêncio; silencioso*	*Distância vertical entre a superfície da água e a parte mais baixa do navio naquele plano, considerando uma embarcação, para um plano transversal de interesse.*
Camisa	*Peça de vestuário, masculino ou feminino, em geral de tecido leve, com mangas curtas ou compridas, e que se veste ordinariamente sobre a pele e vai desde o pescoço até a altura dos quadris, fechada na frente por uma fileira de botões.*	*Componente da bomba de fundo, utilizado no método de produção por bombeio mecânico, consistindo de um tubo que envolve o pistão e que possui superfície interna polida e perfeitamente ajustada ao pistão para que, durante o ciclo de bombeio, seja minimizado o escorregamento de fluido bombeado.*
Campo	*Terreno extenso e plano; terreno plano, extenso, com poucas árvores; campina.*	*Área definida por critérios técnico-administrativos que contém uma ou mais acumulações comerciais conhecidas de petróleo em unidades tectônicas, tais como uma bacia sedimentar, ou em geossinclinais, ou seja, em grande bacia geológica que recebeu a sedimentação de grandes espessuras de sedimentos originadas das áreas adjacentes mais elevadas.*
Coque	*Pancada leve na cabeça com os nós dos dedos; carolo, cascudo.*	*Combustível derivado da aglomeração de carvão e que consiste de matéria mineral e carbono, fundidos juntos.*
DST	*Sigla de Doença Sexualmente Transmissível*	*Drill Stem Test (Teste de Formação)*
Fadiga	*Cansaço resultante de trabalho, físico ou intelectual, intenso; fadigação, pregação.*	*Dano progressivo (acumulativo), localizado e permanente que ocorre no material quando a estrutura está submetida a tensões cíclicas.*
Peixe	*Animais aquáticos, vertebrados, cujos membros são nadadeiras sustentadas por raios ósseos, com esqueleto ósseo ou cartilaginoso, com pele revestida de escamas, cuja respiração é feita por brânquias.*	*Peça metálica deixada dentro de um poço de petróleo que constitua impedimento ao prosseguimento normal das operações de perfuração.*
Reserva	*Ação ou efeito de reservar(-se); reservação.*	*Volumes de petróleo e de gás natural estimados como comercialmente recuperáveis pela aplicação de projetos de desenvolvimento em acumulações conhecidas, a partir de uma determinada data, sob condições definidas.*
Reservatório	*Lugar para armazenar algo; depósito.*	*Configuração geológica dotada de propriedades específicas, armazenadora de petróleo ou gás em subsuperfície.*
Testemunho	*Narração real e circunstanciada que se faz em juízo; declaração, depoimento.*	*Cilindro de rocha cortado durante a perfuração de poços, que varia normalmente de 2 cm a 25 cm de diâmetro e em vários metros no comprimento.*
Unha	*Garra recurva e pontiaguda de alguns animais.*	*Acessório utilizado para calçar estruturas.*

Tabela 1.2: Definições de termos técnicos sem equivalência no dicionário Michaelis

Termo	Definição: Dicionário do Petróleo
BCS	*Bombeamento Centrífugo Submerso*
BOP	*Blowout Preventer (Preventor de Erupção, Obturador de Segurança)*
CLF	*Conector de Linha de Fluxo (Flowline Hub)*
DHPT	*Downhole Pressure&Temperature (Temperatura e Pressão no Fundo do Poço)*
FPSO	*Floating, Production, Storage and Offloading (Unidade Flutuante de Produção, Estocagem e Transferência de Óleo)*
HWCT	*Horizontal Wet Christmas Tree (Árvore de Natal Molhada Horizontal)*
MMBOE	*Milhões de Barris de Óleo Equivalente (Million Barrels of Oil Equivalent)*
SMS	*Segurança, Meio Ambiente e Saúde (Safety, Environment and Health)*
VLCC	*Very Large Crude Carrier (Navio-tanque de Petróleo)*

1.3 Motivação

Avanços recentes na disponibilidade de técnicas de processamento de linguagem natural têm impulsionado um crescente interesse pelo desenvolvimento de pesquisas aplicadas à indústria do petróleo. Entretanto, apesar da forte demanda por soluções que viabilizem um processamento eficiente do vasto volume de dados textuais disponíveis em bases corporativas, paradoxalmente, a oferta de material público em português adequado para o treinamento de modelos de PLN especializados no domínio de O&G é escassa na literatura científica.

Nesse sentido, é relevante destacar como a evolução da área de Inteligência Artificial (IA) somente foi possível a partir das importantes contribuições advindas de iniciativas voltadas a promover bases de dados para treinamento dos algoritmos, como o MNIST (LECUN *et al.*, 1998), PASCAL (EVERINGHAM *et al.*, 2006), CIFAR (KRIZHEVSKY, 2009), SQuAD (RAJPURKAR et al., 2016) e GLUE (WANG et al., 2018). Conforme relatado por DELIPETREV (2020), uma decisiva mudança de paradigma foi estabelecida pelo advento do ImageNet (DENG *et al.*, 2009), cujos autores defenderam que a principal limitação para o progresso da área de IA estaria não apenas no desenvolvimento de novos algoritmos, mas principalmente na disponibilidade de dados adequados para refletir cenários do mundo real. O subsequente sucesso obtido por algoritmos desenvolvidos a partir da base do ImageNet (KRIZHEVSKY *et al.*, 2012; ZEILER e FERGUS, 2014; SIMONYAN e ZISSERMAN, 2014; HE *et al.*, 2016) foram determinantes para moldar o curso da área de IA e contribuíram para impulsionar o intenso crescimento das pesquisas em aprendizagem profunda (RUDER, 2018). Portanto,

dispor de conjuntos de dados públicos e representativos do domínio de O&G é fundamental para fomentar pesquisas nesse domínio e dar suporte ao desenvolvimento de soluções especializadas.

Além disso, a dificuldade de se estabelecer métricas de avaliação adequadas para os modelos representa um desafio adicional para viabilizar a adoção dessas soluções em ambientes de produção industriais. Apesar da popularidade obtida pelos modelos de vetorização, não há consenso na comunidade científica sobre a melhor forma de oferecer uma adequada avaliação de qualidade dessas representações semânticas de linguagem (SCHNABEL *et al.*, 2015; BARAKOV, 2018), principalmente quando aplicados a um domínio específico (NOORALAHZADEH, 2020), considerando que cada abordagem de avaliação objetiva atuar em um aspecto linguístico diferente (WANG *et al.*, 2019). Nesse contexto, o domínio de O&G carece de metodologias de avaliação e de dados anotados que permitam estabelecer métricas de qualidade para os modelos vetoriais especializados.

Portanto, buscando suprir algumas dessas lacunas, a principal motivação para este trabalho consiste em prover alguns dos principais fundamentos estruturantes para viabilizar o desenvolvimento de soluções de PLN especializadas para o domínio de O&G em português, como os corpora de domínio, os modelos vetoriais de palavras e metodologias de avaliação quantitativas e qualitativas. Dessa forma, acreditamos que muitos pesquisadores da indústria e da comunidade acadêmica possam se beneficiar com a disponibilidade desses recursos, seja para estimular o desenvolvimento de novas soluções digitais de inteligência artificial, ou para fomentar novas iniciativas de pesquisa nesse importante domínio da indústria nacional.

Sob essa perspectiva, as motivações para este trabalho podem ser caracterizadas segundo duas óticas principais: *acadêmico-científicas* e *econômico-estratégicas*.

Sob o ponto de vista *acadêmico-científico*, destacam-se:

a. A crescente disponibilidade de técnicas modernas na área de PLN, atendendo à demanda por soluções especializadas que permitam um melhor aproveitamento das informações em formato textual.

b. A oferta de modelos pré-treinados genéricos em português, passíveis de serem reutilizados com técnicas de *transfer learning*. Entretanto, estudos recomendam desenvolver modelos especializados no domínio em que serão utilizados (LAI *et al.*, 2016; DIAZ *et al.*, 2016; PAKHOMOV *et al.*, 2016;

NOORALAHZADEH *et al.*, 2018; WANG *et al.*, 2018, ALSENTZER *et al.*, 2019, TSHITOYAN *et al.*, 2019; GOMES *et al.*, 2021).

c. O vocabulário estritamente técnico do domínio de Óleo e Gás possui especificidades semânticas que o difere do senso comum, motivando o desenvolvimento de representações específicas para contemplar suas particularidades (GOMES et al, 2018, 2021; NOORALAHZADEH *et al.*, 2018; CORDEIRO *et al.*, 2019).

d. A escassez na literatura científica de corpora públicos e modelos de linguagem para o domínio de O&G, para subsidiar o desenvolvimento de soluções de PLN especializadas.

e. O domínio de O&G carece de dados anotados e de uma metodologia de avaliação que permitam estabelecer métricas de qualidade para os modelos.

f. Métodos recentes de processamento de linguagem natural com abordagens de aprendizagem automática, já consolidados no ambiente acadêmico, têm conquistado espaço no ambiente industrial e têm sido amplamente utilizados com sucesso para resolução de problemas concretos em cenários reais (ITOO *et al.*, 2016; BLINSTON e BLONDELLE, 2017; NOORALAHZADEH et al., 2018; CORDEIRO *et al.*, 2019; TSHITOYAN *et al.*, 2019; KHALIBRI et al., 2019; MISHRA e SHARMA, 2019; KALYAN e SANGEETHA, 2020).

Sob o ponto de vista *econômico-estratégico*, podemos citar:

a. O pré-sal brasileiro representa uma importante fronteira exploratória para a indústria de O&G, cujo enorme potencial exploratório tem atraído interesse para investimentos em projetos de Exploração e Produção (CLAVIJO *et al.*, 2019). Além disso, a maior parte de sua documentação está em português.

b. O português é um dos idiomas com o maior número de falantes nativos no mundo. Mas, apesar da importância do idioma para a indústria de O&G mundial, a disponibilidade de corpora e de modelos vetoriais públicos especializados nesse domínio é muito escassa na literatura científica.

c. As companhias – em especial a Petrobras – dispõem de um farto volume de informações não estruturadas, principalmente no formato texto. Há um

manancial de informações potencialmente valiosas não utilizadas em seu completo potencial. Há grande demanda para o desenvolvimento de soluções tecnológicas que permitam lidar com dados textuais de maneira eficiente a partir de algoritmos computacionais.

d. A Petrobras, em particular pelo seu histórico de atuação e especialidade técnica, posiciona-se como referência na indústria nacional e revela-se como uma das instituições com maior potencial para contribuir com esta pesquisa, por estar apta a prover dados representativos no domínio de O&G em português em quantidade e qualidade suficientes para viabilizar iniciativas de composição de corpora e treinamento dos modelos vetoriais especializados, além de oferecer cenários reais de uso para aplicação prática dos algoritmos.

e. Por fim, a atuação profissional do autor desta tese no Centro de Pesquisas da Petrobras contribui para uma interlocução acadêmico-industrial, com recíproco interesse em promover o avanço da ciência em linhas de pesquisa de PLN aplicadas ao setor de O&G.

1.4 Objetivos e Hipótese

O objetivo geral deste trabalho é desenvolver, avaliar e disponibilizar o **`PetroVec`**: um conjunto de modelos de vetorização de palavras especializados no domínio de Óleo e Gás em português, treinados a partir de *corpora* específicos, compostos por documentos técnicos e acadêmicos publicados em português por instituições de referência atuantes nesse setor. Adicionalmente, objetiva-se viabilizar uma extensiva validação dos modelos através de metodologias de avaliação intrínseca e extrínseca, além de análises qualitativas, estabelecendo métricas comparativas dos modelos especializados em relação a um modelo público de contexto geral em português de referência na literatura científica.

A hipótese central para este trabalho é:

(H.1) *Modelos de vetorização de palavras especializados no domínio de Óleo e Gás em português, treinados a partir de corpora específicos, são capazes de melhorar a qualidade de suas representações semânticas e o desempenho quando aplicados em tarefas de PLN específicas do domínio, comparativamente em relação a um modelo pré-treinado a partir de corpora de contexto genérico?*

A fim de endereçar essa hipótese, faz-se necessário suprir alguns insumos fundamentais para permitir o treinamento e avaliação dos modelos, cujos desdobramentos encontram-se descritos nos seguintes objetivos específicos:

(i) Coletar, reunir e disponibilizar para a comunidade acadêmica um significativo conjunto de dados textuais (c*orpus*) em português representativo do domínio de O&G, composto por documentos técnicos e acadêmicos publicados por instituições nacionais de referência nessa área de conhecimento.

(ii) Treinar e disponibilizar modelos de vetorização de palavras especializados no domínio de O&G em português, utilizando dois dos principais algoritmos disponíveis e amplamente utilizados: Word2vec e FastText.

(iii) Avaliar os modelos **`PetroVec`** com metodologia quantitativa baseada em análise intrínseca, atribuindo uma métrica sobre a capacidade dos modelos em capturar propriedades semânticas a partir do corpus, provendo um *dataset* de similaridade semântica de pares de termos anotados por especialistas em geociências.

(iv) Avaliar os modelos **`PetroVec`** com metodologia quantitativa baseada em análise extrínseca, submetendo os modelos vetoriais a uma tarefa específica de Reconhecimento de Entidades Nomeadas (REN) em geociências.

(v) Além das metodologias quantitativas, realizar uma cobertura abrangente de avaliações baseadas em análises qualitativas, contemplando analogias de palavras, coerência de espaço semântico e categorização de conceitos.

(vi) Comparar os resultados das avaliações em relação a um modelo público pré-treinado de contexto geral em português (HARTMANN *et al.*, 2017), e a um modelo especializado em O&G apresentado em um trabalho preliminar deste autor (GOMES *et al.*, 2018), de modo a servirem de referência para estabelecer parâmetros comparativos sobre os resultados observados.

(vii) Compartilhar de forma pública todo o material produzido com a comunidade científica, incluindo todo o código-fonte, modelos vetoriais e corpora.

1.5 Contribuições

As principais contribuições deste trabalho podem ser assim resumidas:

(i) ***Criação do maior corpus de referência atualmente reportado para o domínio de O&G em português:*** coleta, processamento e disponibilização do maior conjunto de dados textual especializado no domínio de Óleo e Gás em português, composto por milhares de documentos técnicos e acadêmicos públicos obtidos a partir de instituições nacionais de referência nesse domínio.

(ii) ***Disponibilização dos modelos vetoriais PetroVec especializados para o domínio de O&G em português:*** treinamento, avaliação e disponibilização de um conjunto de modelos de vetorização de palavras treinados a partir dos corpora especializados de domínio, contemplando os algoritmos Word2vec e FastText. Esses modelos vetoriais especializados representam um dos principais fundamentos para o adequado desempenho de algoritmos de PLN e podem ser utilizados em inúmeras aplicações acadêmicas, comerciais ou industriais relacionadas ao domínio de O&G.

(iii) ***Disponibilização do primeiro dataset anotado para avaliação intrínseca de similaridade semântica para o domínio de O&G em português:*** em colaboração com pesquisadores das Universidades PUC-RS e UFRGS, e com apoio do Centro de Pesquisas e Desenvolvimento da Petrobras, viabilizamos a criação de um *dataset* contendo índices de similaridade semântica para 1500 pares de termos técnicos do domínio, anotado por especialistas em Geociências da indústria e academia, refletindo as particularidades semânticas específicas do domínio. Esse *dataset* permitiu o cálculo de uma métrica de qualidade e viabilizou a realização de uma metodologia de avaliação intrínseca para os modelos vetoriais pré-treinados.

(iv) ***Desenvolvimento de uma metodologia abrangente de avaliações para os modelos especializados PetroVec:*** condução de uma iniciativa pioneira no domínio de O&G em português para viabilizar um detalhado conjunto de avaliações, de forma a garantir a qualidade semântica dos modelos vetoriais especializados. A metodologia de avaliações contempla métricas quantitativas, baseadas em análises intrínsecas e extrínsecas, além de uma

série de análises qualitativas, baseadas em analogias de palavras, coerência de espaço semântico e categorização de conceitos. Os resultados são comparados com modelos públicos pré-treinados de referência acadêmica para ratificar a relevância dos resultados obtidos, complementado por métricas de significância estatística.

(v) ***Disponibilização de todo material gerado neste trabalho de forma pública e aberta:*** todos os artefatos desenvolvidos nesta tese estão disponíveis para a comunidade científica no repositório **Petrolês**[7], incluindo os corpora e os modelos vetoriais pré-treinados **`PetroVec`**, além dos *datasets* anotados e todo o código-fonte desenvolvido para o pré-processamento, treinamento e avaliação dos modelos, disponíveis no repositório público Github[8]. Espera-se que a divulgação desse material possa colaborar para fomentar novas iniciativas de pesquisa em PLN para o setor de O&G.

(vi) ***Disponibilização do Visualizador de Espaço Semântico dos modelos PetroVec[9]:*** criação e publicação de um ambiente para visualização interativa e experimentação dos modelos **`PetroVec`**, disponível no portal Petrolês, que permite realizar análises exploratórias do espaço semântico dos modelos com projeções em 2D e 3D usando PCA e t-SNE, testar operações de similaridade entre termos do vocabulário técnico e explorar suas regiões de vizinhança.

1.6 Publicações

O objeto de estudo principal e os resultados apresentados neste trabalho são descritos no artigo publicado pelo *journal Computers in Industry*:

(i) GOMES, D., CORDEIRO, F., CONSOLI, B., *et al.*, Portuguese word embeddings for the oil and gas industry: Development and evaluation, **Computers in Industry**, https://doi.org/10.1016/j.compind.2020.103347, 2021.

[7] Repositório público de artefatos de PLN para a indústria de O&G em Português. Disponível em: http://petroles.ica.ele.puc-rio.br/
[8] https://github.com/Petroles/Petrovec
[9] Visualizador de Espaço Semântico dos modelos PetroVec: http://petroles.ica.ele.puc-rio.br/projector.html

Uma versão preliminar deste trabalho de pesquisa, com um escopo reduzido e realizado sobre um conjunto de dados significativamente menor, encontra-se publicado no artigo apresentado na Conferência da Rio Oil & Gas:

(ii) GOMES, D., CORDEIRO, F., EVSUKOFF, A. Word Embeddings em português para o domínio específico de óleo e gás. **in: Proceedings of Rio Oil & Gas Expo and Conference,** 2018.

No contexto do tema de pesquisa e diretamente relacionados ao objeto de estudo principal, os seguintes artigos também foram publicados no decorrer do curso de doutoramento:

(iii) CONSOLI, B., SANTOS, J., GOMES, D. *et al.* Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. **in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC)**. Marseille, França, European Language Resources Association, p. 4625–4630. ISBN 979-10-95546-34-4, 2020.

(iv) CORDEIRO, F., GOMES, D., GOMES, F., *et al.* Technology Intelligence Analysis Based on Document Embedding Techniques for Oil and Gas, **Offshore Technology Conference**. doi:10.4043/29707-MS, 2019.

(v) KRUEL, R., RODRIGUES, M., CORDEIRO, F., *et al*. Busca semântica (tipo Google) para recuperação mais inteligente de informação de Reservatórios e Exploração [10] . **13o Seminário de Reservas e Reservatórios (SRR 2019)**. Universidade Petrobras, Rio de Janeiro. 2019

Por fim, o seguinte *preprint* foi disponibilizado como produto parcial da etapa de levantamento bibliográfico, para compartilhamento do conhecimento com a indústria e comunidade científica atuantes na disciplina de PLN e suas aplicações em O&G:

(vi) GOMES, D., EVSUKOFF, A. **Processamento de linguagem natural em português e aprendizagem profunda para o domínio de Óleo e Gás**. arXiv:1908.01674, 2019.

[10] Artigo premiado como **melhor trabalho técnico** no 13º Seminário de Reservas e Reservatórios (SRR 2019)

1.7 Organização da Tese

Esta tese está organizada em 5 capítulos, conforme abaixo descritos:

- **Capítulo 1 – Introdução:** apresenta os conceitos gerais da tese, as motivações que originaram o projeto de pesquisa, a descrição do problema e os objetivos da pesquisa, além das publicações realizadas pelo autor no decorrer do curso de doutoramento;
- **Capítulo 2 – Fundamentação teórica:** apresenta os pressupostos teóricos que orientam a condução da linha de pesquisa, os principais trabalhos relacionados e uma revisão do estado-da-arte das tecnologias e métodos relacionados ao objeto de estudo principal da tese, contextualizando-os ao domínio de O&G e ao idioma português.
- **Capítulo 3 – Corpora de O&G e modelos de linguagem:** apresenta o processo de coleta e construção do corpus especializado de domínio e o treinamento dos modelos vetoriais **`PetroVec`**.
- **Capítulo 4 – Avaliação dos modelos `PetroVec`:** apresenta a descrição das metodologias de avaliação realizada sobre os modelos **`PetroVec`**, contemplando métricas quantitativas de análise intrínseca e extrínseca, além de um conjunto de análises qualitativas.
- **Capítulo 5 – Conclusões e trabalhos futuros:** apresenta as considerações finais sobre a pesquisa e os principais resultados observados, além das perspectivas de evolução da pesquisa que podem ser realizadas em trabalhos futuros.

Capítulo II

Fundamentação Teórica

> "*Increasingly, human intellectual activities will be performed in conjunction with intelligent machines. Our intelligence is what makes us human, and AI is an extension of that quality.*"
>
> Yann LeCun

A fim de contextualizar a disciplina de processamento de linguagem natural e suas particularidades para aplicações em domínios específicos, em especial considerando as recentes abordagens com técnicas de aprendizagem profunda, este capítulo apresenta os fundamentos teóricos identificados na revisão de literatura que conceituam a linha de pesquisa.

2.1 Processamento de Linguagem Natural

Nas últimas décadas, a indústria e a comunidade científica presenciaram uma expansão sem precedentes na quantidade de informações capturadas e armazenadas em formatos não estruturados, em particular, documentos em diferentes formatos textuais. Observou-se a evolução de um cenário inicialmente dominado por informações estruturadas, armazenadas em sistemas gerenciadores de bases de dados e *data warehouses*, para a situação atual em que predominam formatos não-estruturados (ITTOO *et al.*, 2016), que representam uma fração de aproximadamente 80% do volume total de dados coletados (BLINSTON e BLONDELLE, 2017).

Dispersa nesses vastos volumes de documentos, existe uma riqueza de informações importantes que, se adequadamente extraídas e processadas, podem fornecer insumos para inúmeros processos decisórios e dar suporte a uma ampla variedade de atividades acadêmicas e industriais (ITTOO et al., 2016). Nesse cenário, algoritmos de Processamento de Linguagem Natural (PLN) apresentam-se com um promissor potencial para viabilizar essas tarefas. Existe uma forte demanda pelo desenvolvimento de técnicas que permitam fazer uso mais adequado do imenso potencial econômico, acadêmico e estratégico das informações armazenadas nesses repositórios.

A disciplina de PLN é um ramo da área de conhecimento em inteligência artificial e engloba um conjunto de técnicas que visam habilitar algoritmos computacionais com a capacidade de processar automaticamente textos escritos em linguagem humana. Objetiva, assim, resolver aspectos inerentemente linguísticos como estrutura sintática, desambiguação de palavras e compreender o escopo semântico de uma sentença (MANNING e SCHÜTZE, 1999). Algoritmos de PLN têm sido amplamente utilizados com sucesso em diversas aplicações acadêmicas e industriais, como tradução automática, reconhecimento de entidades nomeadas, classificação de texto, análise de sentimentos, sistemas de perguntas e respostas, extração da informação, busca semântica, sumarização, entre outros, conforme apresentado em detalhes na Seção 2.1.1.

Oportunamente, no contexto de linguística, cabe definir alguns conceitos que serão comumente referenciados no decorrer desta tese: fonologia é o ramo destinado ao estudo de sons na linguagem; morfologia dedica-se à estrutura interna de formação das palavras, a partir da composição de suas unidades básicas (morfemas); sintaxe se refere à

relação estrutural das palavras na formação das sentenças; enquanto semântica se refere à compreensão e sentido do texto (KHURANA *et al.*, 2018).

Os primeiros estudos em PLN iniciaram na década de 1950, resultantes da interseção dos primeiros esforços nos campos de inteligência artificial e linguística (KHURANA *et al.*, 2017). Originalmente, havia distinções conceituais entre PLN e a área de recuperação da informação (*information retrieval*) (MANNING *et al.*, 2008). No entanto, a evolução dessas disciplinas de alguma forma convergiu ao longo do tempo, de maneira que a área de PLN demanda do pesquisador conhecimentos em uma ampla diversidade de especializações (NADKARNI *et al.*, 2011), como linguística computacional, estatística e aprendizagem automática. As abordagens iniciais em PLN tradicionalmente se fundamentavam em análise segmentada (*parsing*) e em bases de regras manualmente elaboradas. No entanto, a natureza essencialmente ambígua da forma de expressão da linguagem e a necessidade de identificar, em uma última instância, a semântica das palavras, ultrapassam a capacidade humana em elaborar bases de regras complexas com uma cobertura adequada (ITTOO *et al.*, 2016). Essa necessidade deu origem à ascensão de novas abordagens com aprendizagem automática, focadas na identificação de padrões a partir de dados de exemplo e fortemente baseadas em métodos estatísticos. Nesse cenário, a área de PLN evoluiu com a aplicação de técnicas como *Support Vector Machines* (SVM), *Hidden Markov Models* (HMM) e *Conditional Random Fields* (CRF) (ITTOO *et al.*, 2016). Dessa forma, durante muito tempo essa área foi dominada por abordagens de aprendizagem automática alimentados por vetores esparsos com muitas dimensões (*one-hot vectors*). No entanto, essas arquiteturas se deparavam com os altos custos computacionais para o processamento desses vetores, e comumente esbarravam com a chamada "maldição da dimensionalidade" (*curse of dimensionality*) (BENGIO *et al.*, 2003).

Mais recentemente, a área de PLN conquistou significativos avanços, especialmente em função da grande disponibilidade de dados textuais, da evolução nas pesquisas em disciplinas linguísticas, incrementos na capacidade computacional disponível e, especialmente, pelo desenvolvimento de novos métodos de aprendizagem automática (HIRSCHBERG e MANNING, 2015). Uma ampla proposição de técnicas viabilizou a adoção de modelos neurais não-lineares baseados em vetores densos como dados de entrada, em substituição aos vetores esparsos (GOLDBERG, 2016). Uma das principais contribuições nessa área pode ser atribuída ao uso de representações

distribuídas de palavras (*distributed word representations*), que se caracterizam por representar palavras ou conceitos através de vetores densos de valores reais com um número de dimensões predefinido (MANNING, 2015). Estas técnicas capacitaram os algoritmos de aprendizagem automática a obter representações das palavras capturando propriedades linguísticas e semânticas a partir do corpus, permitindo estabelecer complexas relações de similaridade e, assim, dar suporte a uma maior capacidade de generalização dos algoritmos, contribuindo para promover o exponencial avanço da área de PLN (YOUNG *et al.*, 2018).

2.1.1 Principais aplicações

No contexto das diversas técnicas relacionadas a problemas de PLN, cabe destacar algumas das aplicações mais comuns nessa área de conhecimento, como tradução automática, reconhecimento de entidades nomeadas, classificação automática de textos, análise de sentimentos, sistemas de perguntas e respostas, extração da informação, sumarização automática e busca semântica, conforme apresentado nos parágrafos a seguir.

A área de tradução automática concentra uma das primeiras iniciativas na linha de pesquisa em PLN, tendo mantido sua relevância até os dias atuais, especialmente considerando a intensa necessidade de comunicação promovida pela globalização das relações e pelas plataformas de redes sociais. Uma das mais importantes contribuições recentemente publicadas nessa área foi o uso de mecanismos de atenção com a arquitetura *Transformer* (VASWANI *et al.*, 2017). Uma detalhada pesquisa sobre o estado-da-arte em *word embeddings* multi-idiomas é apresentada por RUDER *et al.* (2018).

O Reconhecimento de Entidades Nomeadas (REN, *Named Entity Recognition*) objetiva identificar entidades (nomes próprios) contidas em um texto. Algumas das principais categorias tradicionalmente consideradas para as entidades são: PESSOA, LOCAL, ORGANIZAÇÃO e TEMPO (LAMPLE *et al.*, 2016), enquanto domínios específicos podem definir seus próprios conjuntos de entidades. YADAV e BETHARD (2018) apresentam uma revisão sobre REN baseados em modelos de aprendizagem automática.

A classificação automática de texto corresponde a uma das mais fundamentais tarefas em PLN, e objetiva atribuir uma ou mais categorias pré-estabelecidas para um determinado texto. Está diretamente relacionada com outras aplicações, como detecção de *spam*, análise de sentimentos e identificação de idioma. Recentemente, abordagens de

aprendizagem profunda vêm conquistado resultados promissores nesse tipo de aplicação, notadamente em função de sua capacidade de abstrair modelos complexos e as relações não-lineares inerentes a esses dados. KOWSARI *et al.* (2019) apresentam uma ampla pesquisa sobre o histórico e o estado-da-arte nessa área de aplicação.

A análise de sentimentos visa aplicar algoritmos computacionais para classificar a polaridade de opiniões relacionadas a uma afirmação, normalmente na forma binária (POSITIVA ou NEGATIVA) ou ternária (POSITIVA, NEUTRA ou NEGATIVA). Essas técnicas têm sido amplamente empregadas em contextos comerciais e industriais. ZHANG *et al.* (2018) apresentam um detalhado estudo sobre abordagens de aprendizagem profunda para problemas de análise de sentimento.

Sistemas de perguntas e respostas (*question answering*) objetivam analisar uma pergunta formulada em linguagem humana e determinar sua resposta. Comumente atuam em um domínio restrito, de maneira que os sistemas de PLN podem explorar o conteúdo de *corpora* do domínio específico na construção de suas bases de conhecimento. O *dataset* academicamente mais relevante nesta área é o *Stanford Question Answering Dataset* (SQuAD)[11] (RAJPURKAR *et al.*, 2016). Há uma estreita relação desta área com a construção de assistentes virtuais para atendimento automatizado (*dialogue systems*), com enorme interesse comercial e industrial nesse tipo de aplicação. Uma revisão sobre diferentes técnicas utilizadas nessa área é apresentada por YOUNG *et al.* (2018).

Extração da informação (*information extraction*) busca extrair informação relevante e estruturada a partir de conteúdos textuais, comumente restritos a um determinado domínio. Faz uso de outras técnicas como reconhecimento de entidades nomeadas, resolução de correferências e extração de relações, entre outros. NIKLAUS *et al.* (2018) apresentam uma visão sobre o estado-da-arte nessa área.

Algoritmos de sumarização automática objetivam reconhecer em um texto os seus trechos mais significativos e o que pode ser descartado, no intuito de compor um resumo legível do seu conteúdo mais relevante. Há essencialmente duas estratégias principais para sumarização: baseadas em extração, cujo conteúdo final é composto por trechos selecionados e extraídos a partir do texto original, sem sofrer modificações; e baseadas em sumarização abstrativa, que consistem em construir representações semânticas internas para o conteúdo original, e então utilizá-las para construir um texto

[11] *The Stanford Question Answering Dataset*. https://rajpurkar.github.io/SQuAD-explorer/

mais próximo da linguagem humana. Um estudo sobre essa área de aplicação é apresentado por ALLAHYARI *et al.* (2017)

Busca semântica (*semantic search*) consiste no uso de mecanismos mais inteligentes para melhor compreender a intenção do usuário ao formular seus critérios de pesquisa por determinada informação. Objetiva considerar o significado dos parâmetros de busca (*query*), indo além de uma simples consulta por palavras-chave indexadas, buscando agregar as diferentes variações de representação incluídas em um mesmo espectro semântico de um determinado conceito. BAST *et al.* (2016) apresentam uma extensa revisão a respeito de busca semântica em textos e bases de conhecimento.

2.1.2 Representações Distribuídas de Palavras

Antes de analisar mais detalhadamente arquiteturas neurais de PLN, é importante compreender a estrutura fundamental utilizada como unidade de entrada para esses algoritmos. Em linguagem natural, podemos considerar a palavra como sendo essencialmente uma unidade básica de significado. Portanto, é necessário estabelecer representações matemáticas adequadas para os dados textuais de entrada, de forma que seja possível processá-los em algoritmos de aprendizagem automática. MANNING (2015) destaca a importância das técnicas de representação para algoritmos de inteligência artificial, que precisam ser capazes de inferir o significado de um conceito ou expressão a partir de suas menores partes constituintes (as palavras).

Inicialmente, abordagens clássicas de vetorização utilizavam a técnica conhecida como *one-hot encoding*, que consiste em utilizar representações categóricas baseadas em vetores esparsos, de forma que cada propriedade (*feature*) é representada por uma dimensão desse vetor (GOLDBERG, 2016). Ou seja, cada termo é representado como um vetor $\mathbb{R}^{|V|\times 1}$ preenchido por zeros (onde $|V|$ é o tamanho do vocabulário), contendo somente uma posição com o valor 1 no índice correspondente à palavra no vocabulário. Outras abordagens tradicionalmente utilizadas em problemas de recuperação da informação consideram técnicas baseadas em TF-IDF (frequência do termo - frequência inversa nos documentos), que objetiva estabelecer representações vetoriais para as sentenças considerando a relevância das palavras em função de sua ocorrência relativa em uma coleção de documento (MANNING, 2008). Porém, os vetores TF-IDF também são esparsos e com alta dimensionalidade, além de não permitirem capturar nuances de representação semânticas das palavras. Além disso, em PLN os vocabulários

de referência podem ser muito grandes (ocasionalmente alcançando milhões de palavras[12]), cujo problema da dimensionalidade pode tornar o processamento desses vetores computacionalmente ineficiente.

Nesse sentido, para que os algoritmos de PLN consigam uma capacidade de generalização adequada, é desejável capturar certas noções de similaridade entre as palavras. Como a técnica *one-hot* representa cada termo como uma entidade totalmente independente, não há nenhuma relação entre os seus vetores. Ou seja, quaisquer pares de palavras terão seus vetores sempre ortogonais e equidistantes entre si (2.1), não guardando, portanto, quaisquer informações relativas sobre seus significados.

$$(v^{óleo})^T v^{gás} = 0 = (v^{cachorro})^T v^{cão} \qquad (2.1)$$

Portanto, um dos principais saltos obtidos na transição dos modelos lineares para os modelos neurais diz respeito à substituição das representações esparsas (*one-hot*), em que cada termo é representado por uma dimensão do vetor, por representações densas de vetores contínuos com dimensões predefinidas (GOLDBERG, 2016). Ou seja, cada propriedade semântica latente é embutida em um espaço vetorial *n*-dimensional e mapeada como um vetor nesse espaço.

Representações distribuídas de palavras (*distributed word representations*), modelos de vetorização de palavras ou *word embeddings* são representações matemáticas obtidas por algoritmos de aprendizagem automática que consistem em mapear cada palavra de um vocabulário para um vetor denso n-dimensional (BENGIO *et al.*, 2003), induzidas a partir do processamento de grandes conjuntos de dados textuais (*corpora*), de forma que termos similares tendem a ter seus vetores posicionados em uma mesma região de vizinhança no espaço vetorial criado (TURNEY e PANTEL, 2010). Essas técnicas são baseadas na hipótese distribucional (HARRIS, 1954; SAHLGREN, 2008), que consideram que palavras com significados semelhantes tendem a aparecer em um mesmo contexto, e são capazes de capturar características essenciais de linguagem como sintaxe e semântica (MIKOLOV *et al.*, 2013a, 2013b; HARTMAN *et al.*, 2017), com grande capacidade de generalização (GOLDBERG, 2016). Portanto, relações de similaridade entre diferentes termos podem ser inferidas a partir do cálculo da distância entre seus

[12] O vocabulário de modelos word2vec pré-treinados disponibilizados pelo Google (https://code.google.com/archive/p/word2vec/) pode alcançar entre 10^5 a 10^7 termos. Fonte: (https://www.tensorflow.org/tutorials/text/word2vec)

vetores, utilizando métricas como a similaridade cosseno (MIKOLOV *et al.*, 2013a, JURAFSKY e MARTIN, 2020), que calcula a similaridade entre dois termos considerando o ângulo entre seus vetores (Equação(2.2).

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \qquad (2.2)$$

Além da evidente melhoria na eficiência computacional de processamento, uma das principais vantagens das representações distribuídas está na capacidade de capturar diversas relações de similaridade entre as palavras nas dimensões embutidas dos vetores, permitindo uma melhor generalização dos modelos de aprendizagem automática. Essa propriedade de generalização dos modelos vetoriais, ao serem integradas em arquiteturas baseadas em redes neurais, foi comprovadamente decisiva para o atingimento de resultados no estado-da-arte em diversas aplicações de PLN (SCHNABEL *et al.*, 2015; LAI *et al.*, 2016; GOLDBERG, 2016; HARTMAN *et al.*, 2017; KHURANA *et al.*, 2017*;* CAMACHO-COLLADOS e PILLEVAR, 2018; YOUNG *et al.*, 2018; PADARIAN e FUENTES, 2019; TSHITOYAN *et al., 2019*).

As pesquisas pioneiras em aprendizagem automática para representações distribuídas de palavras remetem a RUMELHART *et al.* (1986). A década de 1990 é marcada por importantes contribuições nessa área (ELMAN, 1991), que estabeleceram as fundações nesse campo de pesquisa e que evoluiu com o advento das técnicas de *Latent Semantic Analisys* (LSA) e modelos de linguagem. BENGIO *et al.* (2003) apresentaram uma arquitetura baseada em rede neural artificial para modelos de linguagem utilizando *word embeddings*. Outras técnicas seguiram como adaptações das propostas anteriores, que levaram à criação de disciplinas de modelagem de tópicos (*topic modelling*), baseadas em técnicas como *Latend Dirichlet Allocation* (BLEI, NG e JORDAN, 2003). COLLOBERT e WESTON (2008) apresentaram o primeiro estudo demonstrando a utilidade prática de vetores de palavras pré-treinados aplicados a tarefas de PLN.

Mais recentemente, um dos principais responsáveis pela popularização de WE em aplicações de PLN pode ser atribuído ao advento do Word2vec (MIKOLOV *et al.*, 2013a, 2013b), que apresentaram métodos de vetorização baseados em redes neurais

computacionalmente eficientes para permitir o treinamentos dos vetores em grande escala, além de *datasets* e métricas de avaliação que evidenciaram sua grande capacidade de generalização ao capturar propriedades sintáticas e semânticas a partir do corpus de treinamento. O Word2vec propõe duas diferentes arquiteturas: *continuous bag-of-words* (CBOW) e *skipgram* (Figura 2.1). O treinamento do modelo CBOW consiste em predizer a palavra central a partir do conjunto de palavras vizinhas dentro de uma janela de contexto de tamanho *k*. Ou seja, o treinamento objetiva maximizar a probabilidade de observar a palavra atual, considerando como entrada as palavras de contexto. O treinamento do *skipgram*, por sua vez, realiza o oposto, objetivando predizer a distribuição (probabilidade) das palavras de contexto a partir de uma palavra central de referência. A função objetivo do *skipgram* consiste em minimizar o somatório dos erros de predição entre todas as palavras na camada de saída (MIKOLOV *et al.*, 2013a).

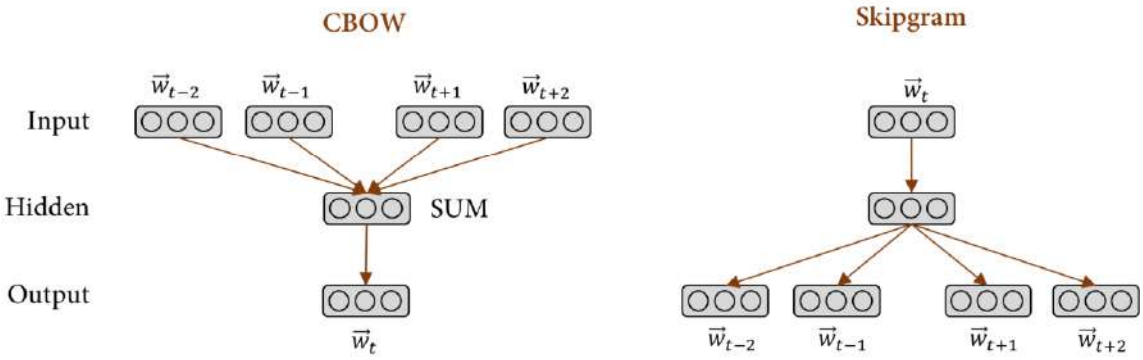


Figura 2.1: Ilustração dos diferentes modelos propostos pelo Word2vec: CBOW e skipgram
Fonte: CAMACHO-COLLADOS e PILLEVAR (2018).

MIKOLOV *et al.* (2013a) evidenciaram, inclusive, que a realização de operações algébricas entre vetores sugere a manutenção das relações semânticas entre os termos (Figura 2.2). Por exemplo, a seguinte operação resulta em uma posição no espaço vetorial que atende à igualdade: $v^{Paris} - v^{France} + v^{Italy} = v^{Rome}$. Os autores também reportam outras relações de analogias semânticas válidas, reproduzidas na Tabela 2.1.

Na sequência, outras importantes técnicas de vetorização de palavras foram propostas. *Global Vectors* (GloVe) (PENNINGTON *et al.*, 2014) é um método de vetorização cujo treinamento é realizado a partir de estatísticas globais de co-ocorrência entre as palavras. BOJANOWSKI *et al.* (2017) apresentaram o FastText, uma variação da arquitetura Word2vec em que os *embeddings* são associados a sequências contíguas de caracteres (*n-grams*), e cada palavra é representada pelo somatório das representações

de seus *n-grams*. Essa característica lhe confere a particular capacidade de atribuir representações inclusive para termos não observados no vocabulário durante o treinamento, além de capturar características morfológicas na formação dos termos. Uma detalhada pesquisa comparando diferentes abordagens de vetorização de palavras é apresentada por LEVY *et al.* (2015). O artigo descreve orientações para otimização dos hiperparâmetros, sugerindo que estes possuem grande impacto na qualidade dos vetores, enquanto não foram observadas grandes variações de desempenho entre os diferentes algoritmos, contrastando com estudos anteriores.

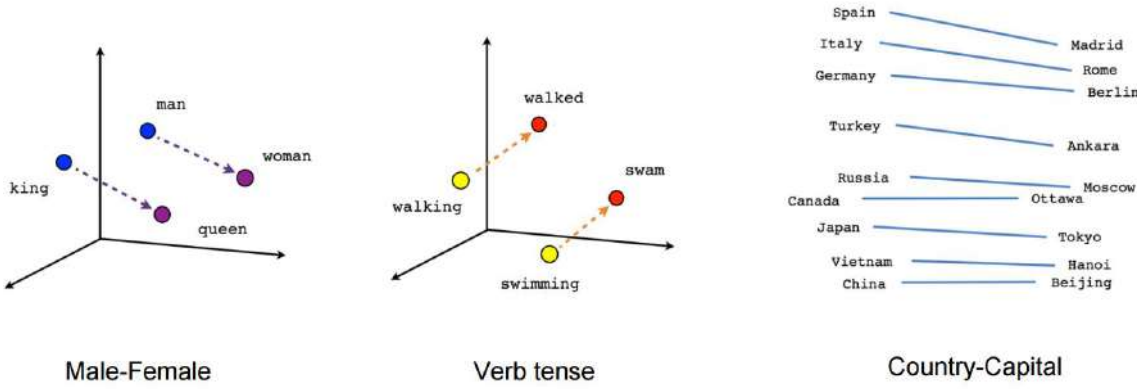


Figura 2.2: Operações vetoriais nos *embeddings* com manutenção de suas relações semânticas
Fonte: Google[13]

Tabela 2.1: Relações de analogias semânticas a partir de operações vetoriais
Fonte: MIKOLOV *et al.* (2013a).

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

No entanto, esses modelos de *word embeddings* computam uma representação estática e fixa para cada palavra, independente do contexto e, portanto, incorrem em algumas limitações ao assumir uma representação única e universal para cada termo. Isso implica em desconsiderar uma característica essencial da linguagem – a polissemia, não

[13] https://www.tensorflow.org/tutorials/word2vec

sendo capazes de representar diferentes significados que uma mesma palavra pode assumir. Essa limitação pode levar a um efeito conhecida como *meaning conflation deficiency* (CAMACHO-COLLADOS e PILEHVAR, 2018), em que os diferentes significados da palavra são projetados em um mesmo ponto do espaço vetorial. Na tentativa de representar unicamente os diferentes significados de um termo, isso pode acarretar o efeito de atrair para seu entorno outras palavras não-semanticamente relacionadas, que estariam por sua vez associadas a regiões representadas pelos diferentes significados do termo (NEELAKANTAN *et al.*, 2014). Portanto, novas abordagens baseadas em representações sensíveis a diferentes contextos foram recentemente propostas, discutidas na Seção 2.1.3.

2.1.3 Representações Contextuais

Para dirimir algumas das limitações das representações estáticas de palavras para lidar com diferentes variações de significado, foram recentemente propostas novas abordagens baseadas em representações sensíveis a diferentes contextos, como ELMO (PETERS *et al.*, 2018), GPT (RADFORD *et al.*, 2018), FLAIR (AKBIK *et al.*, 2018), BERT (DEVLIN *et al*., 2019) e XLNet (YANG *et al*., 2019). Representações contextuais de palavras assumem uma nova perspectiva na atribuição de vetores para cada termo de referência, cujas representações podem variar conforme o contexto em que o termo ocorre. Ou seja, o vetor de uma palavra assume valores dinamicamente diferentes entre o processamento de uma sentença e outra, dependendo de uma janela de contexto. Os modelos contextuais demonstraram ser capazes de capturar não apenas variações semânticas contextuais das palavras, mas também diversas propriedades sintáticas (HEWITT e MANNING, 2019).

Desde o advento das técnicas de *word embeddings*, a abordagem tradicionalmente utilizada para implementar algoritmos de PLN consiste em inicializar a primeira camada das redes neurais com vetores estáticos pré-treinados (como Word2vec e FastText), técnica conhecida como transferência de aprendizado, enquanto o restante da rede é treinado para uma finalidade específica. Nesse sentido, recentes abordagens propõem uma nova estratégia, que consiste em pré-treinar toda a arquitetura de rede profunda e transferir todas as suas camadas, especializando apenas as camadas adicionais de topo para serem reaplicadas em outra tarefa específica (RUDER, 2019). Nesse sentido, esses *embeddings* contextuais de palavras correspondem, de fato, a representações dos

estados internos de uma rede neural profunda, treinada como um modelo de linguagem a partir de um determinado corpus (PILEHVAR e CAMACHO-COLLADOS, 2020).

O *Embedding from Language Model* (ELMo) (PETERS *et al.*, 2018) apresentou uma implementação de representações contextuais de palavras a partir de duas redes LSTM unidirecionais, além de utilizar representações internas baseadas em caracteres, permitindo considerar características morfológicas dos termos e tornando possível a representação de palavras fora do vocabulário de treinamento. Os autores disponibilizaram modelos públicos pré-treinados no repositório do projeto[14], inclusive para o idioma português. O *Universal Language Model Fine-tuning* (ULMFit) (HORWARD e RUDER, 2018) endereça problemas de classificação, sendo ainda possível sua reutilização em outras tarefas utilizando *transfer learning*. O *Generative Pre-Trained Transformer* (GPT) (RADFORD *et al.*, 2018) foi uma das primeiras implementações a substituir as redes LSTM pela arquitetura Transformer (VASWANI *et al.*, 2017), obtendo significativos ganhos de desempenho. Entretanto, assim como ELMo, o GPT baseia-se no uso de modelos de linguagem (*language models*) unidirecionais em sua etapa de pré-treinamento – segundo MANNING e SOCHER (2017), modelos de linguagem objetivam computar a probabilidade de ocorrência de um conjunto de palavras em uma sequência particular. Entretanto, uma das limitações dessa abordagem unidirecional é não considerar simultaneamente o contexto tanto à direita como à esquerda da palavra-alvo, de maneira que a rede computa apenas os tokens anteriormente vistos na sequência de entrada. Nesse sentido, DEVLIN *et al.* (2019) apresentaram o *Bidirectional Encoder Representations from Transformers* (BERT), que estabeleceu novos patamares de estado-da-arte para diferentes *benchmarks* de PLN. Semelhante ao GPT, esse modelo baseia-se na arquitetura Transformer, porém introduz um mecanismo bidirecional que permite considerar o contexto das palavras tanto à direita como à esquerda de forma verdadeiramente bidirecional em todas as camadas da rede. Dessa forma, os modelos obtiveram excelentes resultados em diversas tarefas de PLN, além de demonstrar que aplicações específicas podem especializar os modelos BERT pré-treinados[15] otimizando apenas uma camada adicional no topo da rede. A Figura 2.3 ilustra conceitualmente as diferentes estratégias de pré-treinamento utilizadas pelas arquiteturas

[14] *Deep contextualized word representations*: https://allennlp.org/elmo
[15] https://github.com/google-research/bert#pre-trained-models

BERT, GPT e ELMo. O Transformer e as principais arquiteturas utilizadas por algoritmos de PLN são abordados em detalhes na Seção 2.1.4.

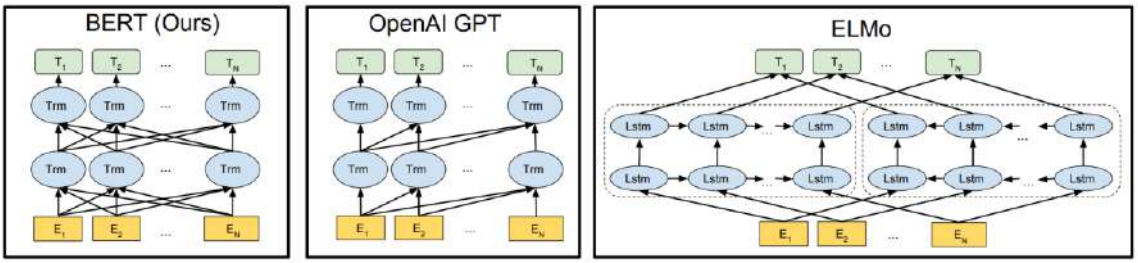


Figura 2.3: Diferenças nas estratégias de pré-treinamento entre as diferentes arquiteturas. BERT utiliza a arquitetura Transformer bidirecional. ELMo utiliza a concatenação de duas redes LSTM unidirecionais independentes, enquanto o GPT utiliza Transformer unidirecional.
Fonte: DEVLIN *et al.* (2019).

Mais recentemente, houve uma ampla proposição de modelos contextuais contemplando variações da arquitetura Transformer, experimentando ajustes de hiperparâmetros e novas estratégias de pré-treinamento em relação ao modelo BERT, como o RoBERTa (Liu *et al.*, 2019), DistilBERT (SANH *et al.*, 2019), ALBERT (ZHENZHONG *et al.*, 2020), GPT-2 (RADFORD *et al.*, 2019) e XLNet (YANG *et al.*, 2019). Uma detalhada revisão, reunindo as principais publicações sobre a arquitetura BERT, é apresentada por ROGERS *et al.* (2020). Adicionalmente, uma ampla pesquisa contemplando algumas das principais variações de BERT e da arquitetura Transformers é apresentada por XIA *et al.* (2020). Para simplificar a construção dos algoritmos computacionais, uma biblioteca contendo uma ampla variedade de implementações de arquiteturas Transformers é apresentada por WOLF *et al.* (2020), disponibilizada publicamente pelos autores no repositório Github[16].

2.1.4 Principais arquiteturas para PLN

Redes neurais recorrentes (*Recurrent Neural Networks*, RNN) (ELMAN, 1990) objetivam executar o processamento dos dados de entrada em sequência. Utilizam um vetor de estados para armazenar uma memória interna (EVSUKOFF, 2020), de maneira que um determinado passo recebe como entrada o resultado do processamento do passo anterior (Figura 2.4). RNNs, uma vez desdobradas, podem ser entendidas como se fossem redes *feed-forward* muito profundas, em que todas as suas camadas compartilham os mesmos pesos (LECUN *et al.*, 2015). Esse compartilhamento dos pesos ao longo de vários passos no processamento sequencial permite o aprendizado de padrões que

[16] Hugginface Transformers: https://github.com/huggingface/transformers

ocorrem em posições distintas na cadeia de entrada, conferindo propriedades que lhe atribuem analogias ao conceito de memória. RNNs se apresentam como uma escolha natural para modelar problemas em PLN, considerando sua capacidade de capturar as características inerentemente sequenciais presentes na linguagem, processadas em cadeias de caracteres, palavras ou sentenças (YOUNG *et al.*, 2018).

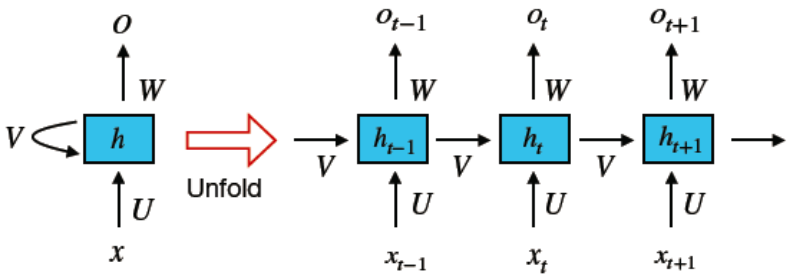


Figura 2.4: Rede Neural Recorrente (RNN) desdobrada em passos iterativos
Fonte: LECUN *et al.* (2015).

As primeiras abordagens neurais para modelos de linguagens foram apresentadas por BENGIO *et al.* (2001) e baseadas em redes *feed-foward*, tendo sido substituídas por arquiteturas baseadas em redes recorrentes (MIKOLOV *et al.*, 2010). No entanto, RNNs apresentam limitações e são difíceis de serem treinadas para identificar dependências em posições muito distantes da cadeia de entrada, em função do problema da explosão ou desaparecimento do gradiente (BENGIO *et al.*, 2013). Portanto, a arquitetura original das RNN comumente empregada em PLN logo evoluiu para adoção de redes *Long Short-Term Memory* (LSTM) (HOCHREITER e SCHMIDHUBER, 1997), por ser mais robusta e escalável ao lidar com o problema do gradiente no decorrer das camadas profundas. Redes LSTM introduzem explicitamente o conceito de células de memória, responsáveis por preservar o gradiente mesmo em cadeias de sequências mais longas (LECUN *et al.*, 2015), além de incluir uma porta adicional para "esquecimento" (GERS *et al.*, 1999). A Figura 2.5 ilustra um diagrama conceitual para uma rede LSTM.

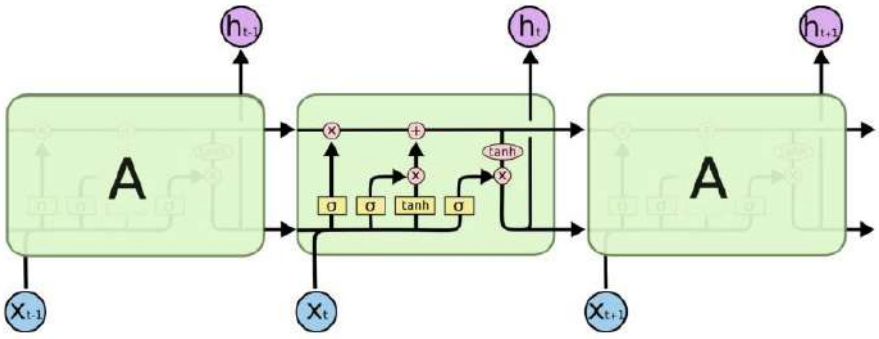


Figura 2.5: Rede *Long Short-Term Memory* (LSTM)
Fonte: OLAH (2015).

Uma arquitetura alternativa para simplificar o complexo funcionamento das células LSTM foi proposta por CHO *et al.* (2014), apresentando a unidade *Gated Recurrent Unit* (GRU), em que uma única porta é responsável por controlar simultaneamente o fator de esquecimento e a decisão de atualizar sua célula interna. JOZEFOWICZ *et al.* (2015) apresentam um detalhado estudo comparativo sobre diferentes arquiteturas de redes recorrentes, incluindo redes LSTM e GRU, explorando suas potencialidades e fraquezas em diferentes cenários.

Redes neurais convolucionais (*Convolutional Neural Networks*, CNN) se popularizaram a partir do enorme sucesso obtido em aplicações na área de visão computacional (KRIZHEVSKY *et al.*, 2012), e rapidamente foram adaptadas para aplicações em processamento de linguagem natural (KIM, 2014). Não obstante, COLLOBERT e WESTON (2008) estão entre os primeiros pesquisadores a registrar a aplicação de CNN em tarefas de PLN. Em arquiteturas CNN, os dados de entrada (caracteres, palavras ou sentenças) são representados como vetores (*embeddings*), que podem ser iniciados de forma aleatória ou transferidos a partir de pré-treinamento. Entretanto, utilizar vetores pré-treinados a partir de *corpora* representativos pode contribuir para melhoria do desempenho (KIM, 2014). Filtros convolucionais (*kernels*) são então aplicados, deslocando-se conforme uma janela de tamanho *n*. Essa operação é repetida em diversas camadas, com variações nos tamanhos dos filtros, intercalando com operações de *max-pooling* para reduzir a dimensão e aumentar o nível de abstração dessas representações. Nesse contexto, em PLN os filtros convolucionais operam iterando ao longo da dimensão temporal representada pelas cadeias de entrada. A Figura 2.6 ilustra tipicamente a utilização de uma CNN em PLN.

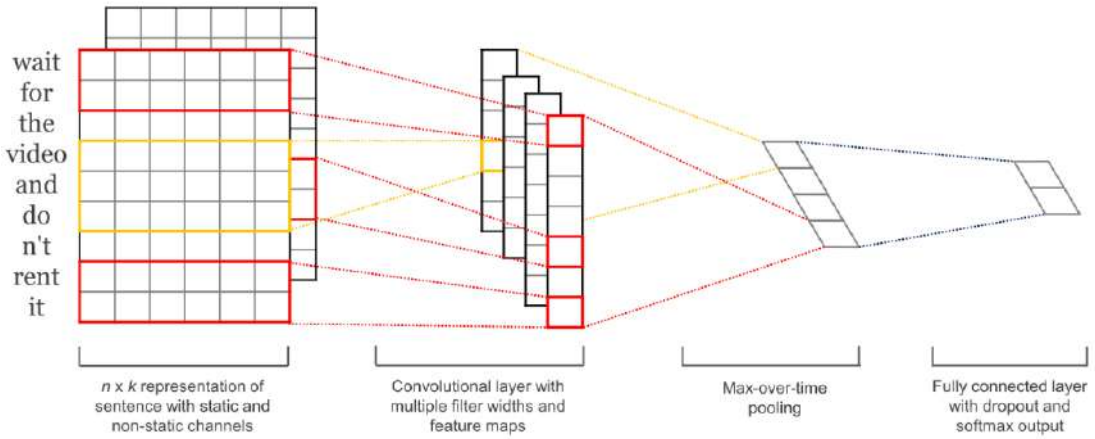


Figura 2.6: Rede neural convolucional (CNN) para processamento de texto
Fonte: KIM (2014).

Uma das vantagens das redes CNN sobre as RNN é sua maior capacidade de paralelização, uma vez que cada passo da convolução depende apenas de seu contexto local, enquanto as RNNs dependem de toda a cadeia de estados anteriores. Nesse sentido, YIN *et al.* (2017) apresentam um estudo comparativo sobre o desempenho de CNN e RNN em diferentes aplicações de PLN, e concluem não haver uma opção universalmente melhor, dependendo, portanto, das condições globais e semânticas envolvidas em cada tarefa específica. Um detalhado levantamento das principais arquiteturas e seus resultados em aplicações de PLN é apresentado por YOUNG *et al.* (2018).

Arquiteturas RNN consideram o processamento do texto subdividindo-o como uma sequência de dados de entrada. Entretanto, é também possível compreender a linguagem segundo uma perspectiva hierárquica, onde as palavras são combinadas de maneira recursiva para compor expressões que, por sua vez, podem ser novamente combinadas em uma ordem mais alta para compor novos conceitos. Essa ideia inspirou a interpretação de sentenças como árvores hierárquicas (SOCHER *et al.*, 2013). Redes neurais recursivas (*Recursive Neural Networks*, RecNN) constroem a representação de suas sentenças segundo uma perspectiva de baixo para cima (*bottom-up*), em contraste com redes recorrentes que consideram como uma sequência da esquerda para a direita ou direita para a esquerda (YOUNG et al, 2018). A representação em uma estrutura de árvore permite que os modelos recursivos possam fazer melhor uso da interpretação sintática da estrutura das sentenças, sendo particularmente útil em situações de negação, por exemplo, em que uma determinada palavra pode produzir efeito direto em todo um segmento da sentença, conforme ilustrado na Figura 2.7.

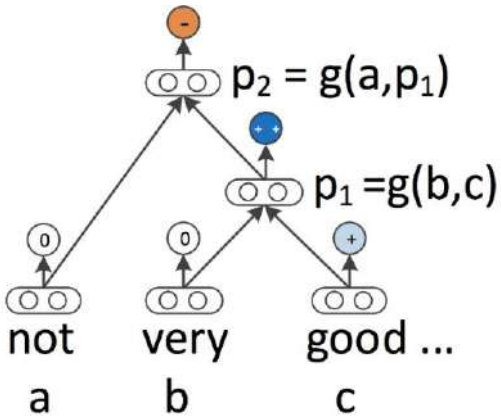


Figura 2.7: Estrutura da rede recursiva aplicada a classificação de sentimentos
Fonte: SOCHER *et al.* (2013).

Recentemente, novas propostas surgiram como evolução das principais arquiteturas anteriormente abordadas. SUTSKEVER *et al.* (2014) apresentaram uma abordagem sequência-a-sequência baseada em arquitetura neural codificador-decodificador (*encoder-decoder*). Uma rede recorrente é usada para codificar a sequência de entrada como uma representação y_n, e esse vetor é usado como entrada auxiliar para uma segunda RNN responsável pela decodificação (GOLDBERG, 2015). Essa arquitetura se mostrou adequada para problemas de tradução automática (*neural machine translation*, NMT) (SUTSKEVER *et al.*, 2014), logo sendo adotada pelo Google (WU *et al.*, 2016) e alcançando resultados de estado-da-arte.

Uma das principais limitações em modelos de sequência-a-sequência está no fato de compactarem todo o processamento da entrada em um único vetor de tamanho fixo (*encoder*), incluindo informações que podem não ser totalmente relevantes no contexto da tarefa atual, ou nos casos em que a entrada é muito longa, que depois será expandida novamente pelo *decoder*. Nesse sentido, o mecanismo de atenção (BAHDANAU *et al.*, 2015) busca endereçar esse problema permitindo que o *decoder* tenha acesso às informações da sequência de entrada. Ou seja, durante a decodificação, além do último estado, o *decoder* também tem acesso a um vetor de contexto calculado em função da sequência de entrada (YOUNG *et al.*, 2018), tornando possível inferir trechos relevantes que merecem mais ou menos atenção no processamento. Uma variação dessa abordagem é conhecida como auto-atenção (*self-attention*) que pode ser utilizada para analisar palavras vizinhas na sentença de entrada a fim de enriquecer as informações contextuais.

Entretanto, um dos principais gargalos nas arquiteturas anteriormente citadas se refere à natureza sequencial de processamento dos dados de entrada, inviabilizando um paralelismo mais eficiente. Para endereçar esse problema, VASWANI *et al.* (2017) propuseram o Transformer, uma arquitetura *encoder-decoder* puramente baseada em mecanismo de atenção, que substituem as recorrências das RNN e convoluções das CNN por múltiplas camadas empilhadas de *self-attention*. Como resultado, o Transformer oferece uma arquitetura altamente paralelizável (Figura 2.8), e que atualmente representa o estado-da-arte em diversas aplicações de PLN. Comparados às RNN que processam o texto de entrada de forma sequencial, Transformers processam todos os dados de entrada em paralelo, tornando-o mais adequados às arquiteturas modernas de GPU e TPUs. Além disso, diferentemente das redes RNN, Transformers podem resolver sequencias de texto

mais longas na cadeia de entrada, relacionando palavras relativamente distantes entre si com maior eficiência (PILEHVAR e CAMACHO-COLLADOS, 2020).

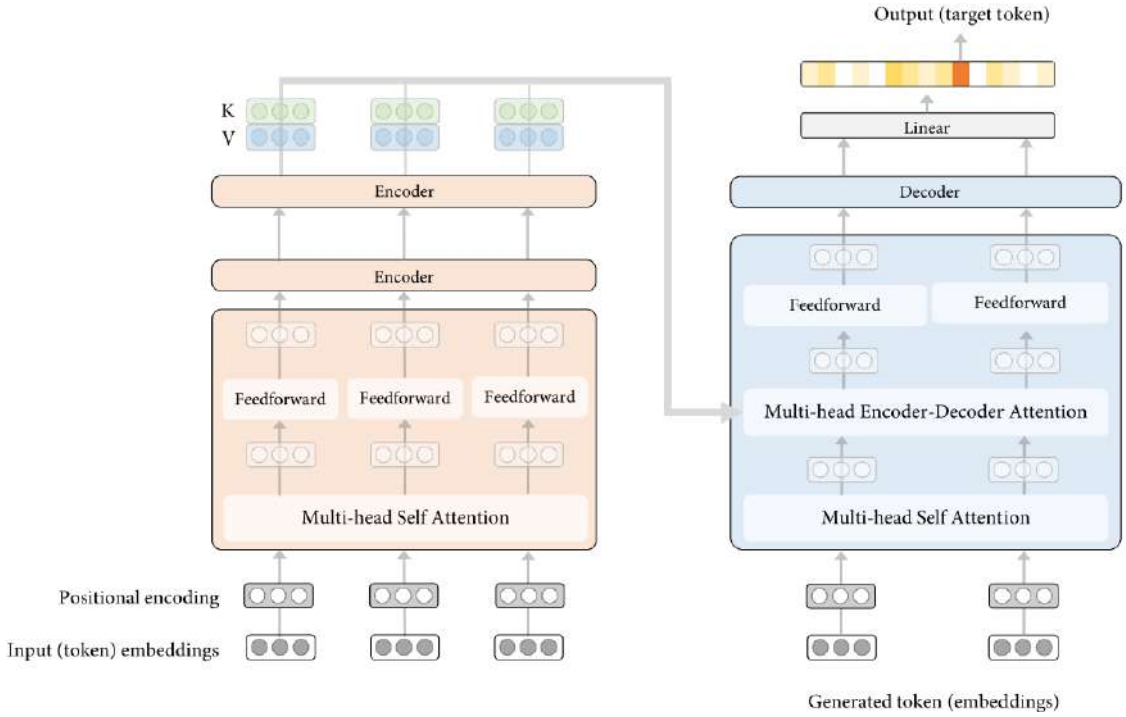


Figura 2.8: Diagrama conceitual com a arquitetura Transformer
Fonte: PILEHVAR e CAMACHO-COLLADOS (2020).

2.1.5 Evolução da complexidade computacional dos modelos contextuais

Apesar dos excelentes resultados recentemente reportados para os modelos contextuais em diversos trabalhos, em contrapartida, há um expressivo aumento nos custos computacionais consequentemente requeridos para o treinamento e inferência desses complexos modelos contextuais baseados em arquiteturas de redes profundas. Em cenários acadêmicos e industriais do mundo real, onde os recursos computacionais são limitados e imensos volumes de dados precisam ser processados em curtos tempos de resposta, a latência representa um importante requisito e deve ser também considerada em complemento à acurácia. Em alguns casos, face a limitações de custo e escopo, a implantação desses modelos pode ser inviável, ou os ganhos de desempenho podem não justificar os altos custos.

Nesse sentido, POLIGNANO *et al.* (2020) reportam que, em situações reais de uso, os ganhos de acurácia podem ser pouco significativos e nem sempre justificam o alto incremento no custo computacional requerido para treinar e executar os complexos modelos contextuais modernos. Os autores explicitamente reforçam a ideia de que os

modelos BERT podem nem sempre ser a melhor escolha para qualquer cenário de aplicação prática, e encorajam análises adicionais sobre o custo-benefício entre latência e acurácia ao optar por tais modelos, recomendando que arquiteturas menos complexas devem ser adotadas nos casos em que os modelos apresentem desempenhos similares.

ARORA *et al.* (2020) conduzem um detalhado estudo comparando o desempenho de modelos de *word embeddings* estáticos e modelos contextuais recentes, analisando seus ganhos de acurácia em relação ao consequente aumento nos requisitos computacionais. Os autores reportam que, em muitos cenários aplicados em ambientes de produção e utilizando dados reais em escala industrial, os modelos de WE estáticos obtiveram resultados altamente competitivos quando treinados com dados em quantidade suficiente, enquanto os modelos contextuais mantiveram melhor desempenho em algumas aplicações que envolvem, especificamente, estruturas textuais mais complexas, ambiguidades e vocabulário desconhecido. Os autores reportam, ainda, que a qualidade e a quantidade dos dados utilizados no treinamento são um fator crucial para determinar a performance relativa entre os modelos estáticos e contextuais, de forma que o desempenho dos modelos estáticos melhora rapidamente à medida em que aumenta a quantidade de dados disponível para treinamento.

Resultados similares foram reportados por RÉ *et al.* (2020) que, motivados pelo excepcional desempenho reportado pelos modelos BERT em *benchmarks* acadêmicos como o GLUE (WANG *et al.*, 2018), realizaram experimentos comparando *embedding* estáticos com modelos BERT aplicados em um sistema em produção na Apple. Os autores relatam, entretanto, que não foram observados ganhos significativos com o uso de modelos BERT para esse cenário, especificamente. Da mesma forma, XIA *et al.* (2020) alertam que o desempenho reportado por esses modelos nos principais benchmarks não necessariamente garante resultados similares quando aplicados em cenários reais de uso com dados do mundo real.

BENDER *et al.* (2021) apresentam uma detalhada discussão a respeito da recente tendência de evolução dos modelos de linguagem, baseadas em arquiteturas cada vez maiores e mais complexas, e analisam seus potenciais riscos e as possíveis direções futuras para viabilidade dessa linha de pesquisa. Os autores reportam uma variedade de riscos e custos relacionados ao tema, como: impactos ambientais, potencialmente associados à maior capacidade computacional demandada para treinamento; custos financeiros elevados, que podem criar limitações para a atuação de determinados grupos

de pesquisa acadêmicos ou instituições de menor porte; custos de oportunidade, que podem motivar pesquisadores a privilegiar idiomas mais representativos; além de eventuais danos sociais como estereótipos, vieses ideológicos e outros impactos ocasionados por uma eventual falta de curadoria dos dados de treinamento, motivados por uma busca indiscriminada por conjuntos de dados maiores e sem uma preocupação adequada com sua qualidade e representatividade.

Nesse contexto, um importante tema tem conquistado espaço na comunidade científica, no sentido de discutir a relação entre os promissores resultados obtidos pelas recentes contribuições advindas dos modelos contextuais baseados em arquiteturas de aprendizagem profunda, e o consequente aumento expressivo dos requisitos computacionais necessários para dar suporte ao treinamento e inferência desses modelos. Observa-se uma rápida evolução na quantidade de parâmetros necessários para implementar arquiteturas cada vez mais complexas das redes neurais profundas (Figura 2.9 e Figura 2.10). Conforme apresentado por MANNING (2020) e reproduzido na Figura 2.11 e Figura 2.12, as novas gerações desses modelos rapidamente superam em complexidade as anteriores em curtos intervalos de tempo, em contrapartida obtendo ganhos de acurácia relativos cada vez menores e que não seguem na mesma proporção do aumento dos custos computacionais.

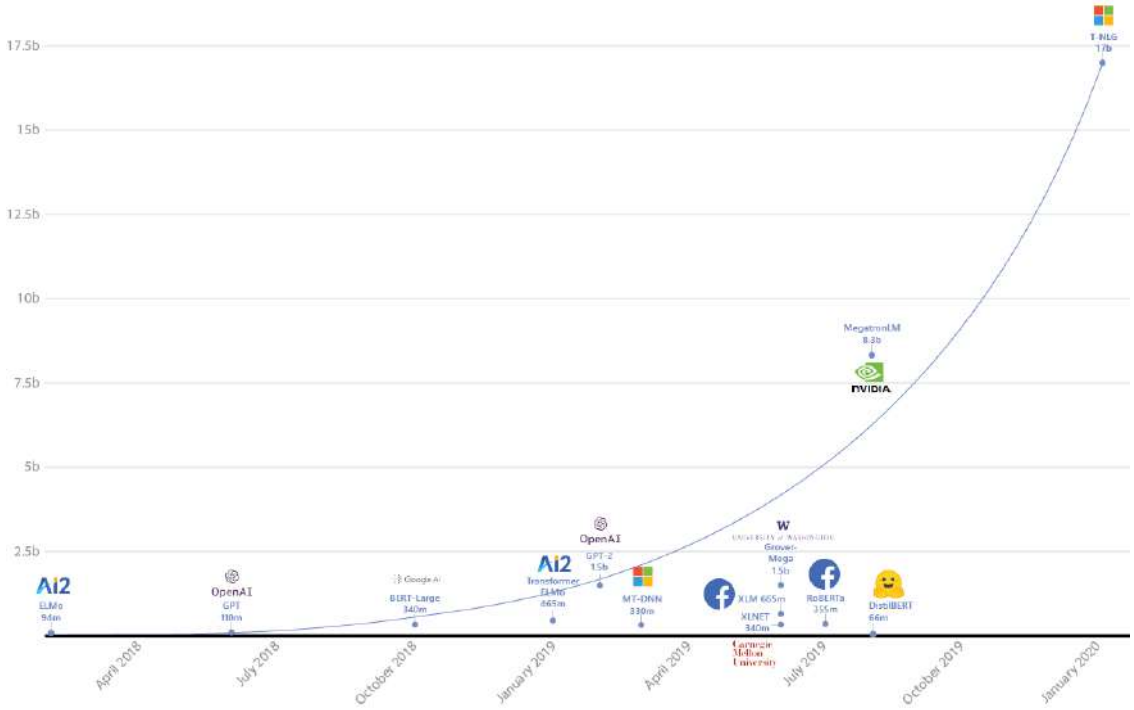


Figura 2.9: Evolução da complexidade arquitetural dos modelos de PLN (número de parâmetros da rede), até janeiro/2020.
Fonte: MICROSOFT (2020).

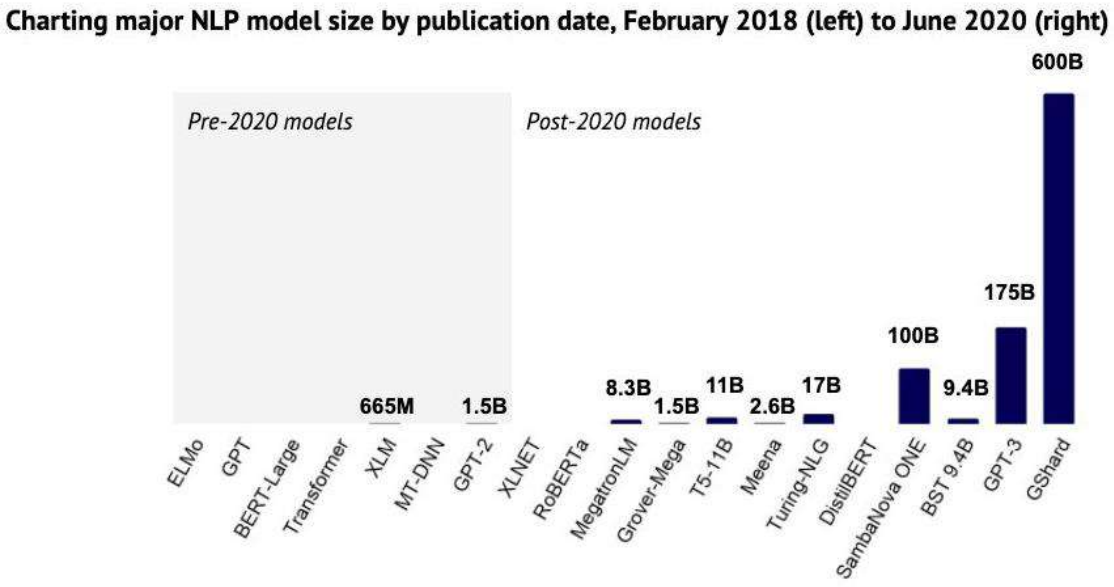


Figura 2.10: Evolução da complexidade arquitetural dos modelos de PLN (número de parâmetros da rede), até a publicação do GShard (LEPIKHIN *et al.*, 2020) em junho/2020.
Fonte: BENAICH e HOGARTH (2020).

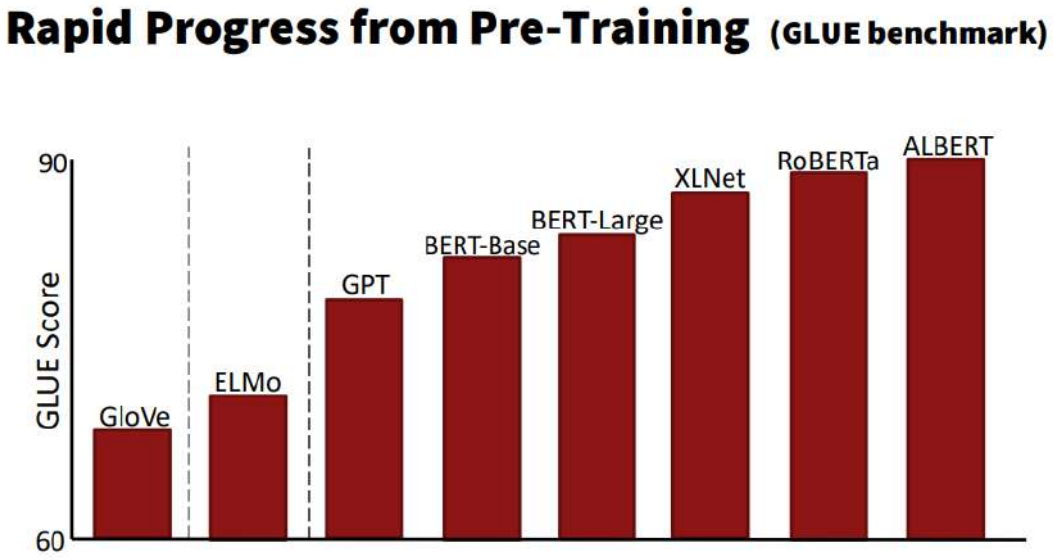


Figura 2.11: Incrementos de acurácia entre as gerações de modelos contextuais, que não acompanham na mesma proporção o aumento dos custos computacionais
Fonte: MANNING (2020).

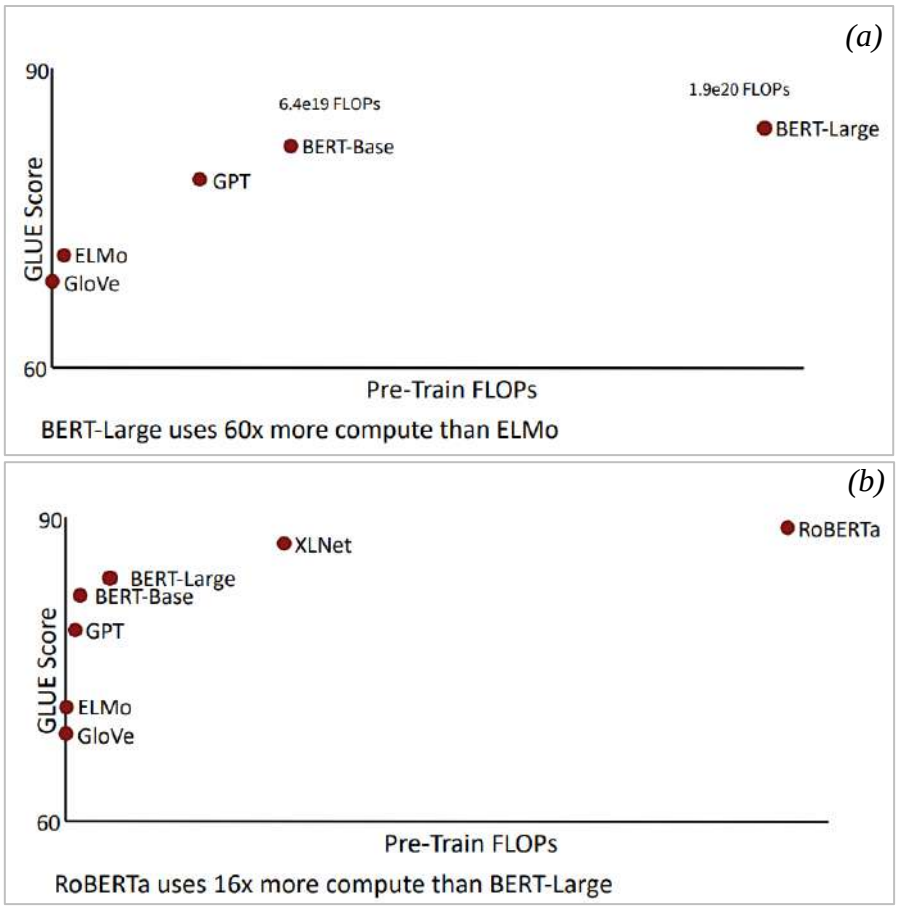


Figura 2.12: Aumento dos custos computacionais considerando as recentes gerações de modelos contextuais: BERT demanda 60x mais computação que o ELMO *(a)*, RoBERTa demanda 16x mais computação que o BERT *(b)*.
Fonte: MANNING (2020).

Portanto, diversos pesquisadores defendem a importância de ampliar as discussões sobre o *tradeoff* entre ganhos de acurácia e o aumento do custo computacional, e incentivam o desenvolvimento de novos métodos para viabilizar a implementação desses complexos modelos de uma forma mais eficiente e sustentável[17] (STRUBELL *et al.*, 2019; SCHWARTZ *et al.*, 2020; MOOSAVI *et al.*, 2020). Um exemplo dessa premissa é o ELECTRA (CLARK *et al.*, 2020), que objetiva oferecer um modelo com uma implementação mais eficiente e capaz de obter resultados similares com menores custos computacionais (Figura 2.13). Uma recente publicação que pode sinalizar uma possível tendência para os modelos baseados em Transformer é apresentada pela equipe do Google Research (KITAEV *et al.*, 2020), que propõe uma arquitetura computacionalmente mais leve e escalável chamada Reformer, capaz de lidar com janelas

[17] Em novembro/2020 ocorreu a primeira edição do evento *SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, com objetivo de encorajar o desenvolvimento de modelos de PLN computacionalmente mais eficientes e arquiteturalmente mais simples: https://www.aclweb.org/anthology/2020.sustainlp-1.0/. Acesso em 12/01/2021.

de contexto de até um milhão de palavras e com uso eficiente de memória. Em seguida, uma equipe do Facebook AI disponibilizou o Linformer (WANG *et al.*, 2020), uma nova arquitetura baseada em Transformer mais escalável para sequências longas. Os autores reportam que o custo computacional de modelos Transformer cresce em escala geométrica à medida que aumenta o tamanho da entrada, enquanto com o Linformer esse crescimento ocorre em escala linear (conforme ilustrado na Figura 2.14), o que lhe permite o treinamento com sequências mais longas, além de viabilizar a aplicação desses modelos em imagens mais facilmente.

Portanto, assim ratificando as motivações e o escopo desta tese, modelos estáticos de *word embeddings* tais como o **`PetroVec`** são mais leves e rápidos para o treinamento e inferência, representando uma alternativa muito eficiente a uma fração do custo computacional quando comparados aos modelos contextuais. Além disso, esses modelos de WE têm sido amplamente utilizados com sucesso em vários cenários reais de uso, tanto em ambientes industriais como acadêmicos, obtendo excelentes resultados em diversas aplicações de PLN (ITOO *et al.*, 2016; BLINSTON e BLONDELLE, 2017; YOUNG *et al.*, 2018; NOORALAHZADEH *et al.*, 2018; CORDEIRO *et al.*, 2019; TSHITOYAN *et al.*, 2019; KHABIRI *et al.*, 2019; PADARIAN *et al.*, 2019; KALYAN e SANGEETHA, 2020; GOMES *et al.*, 2021).

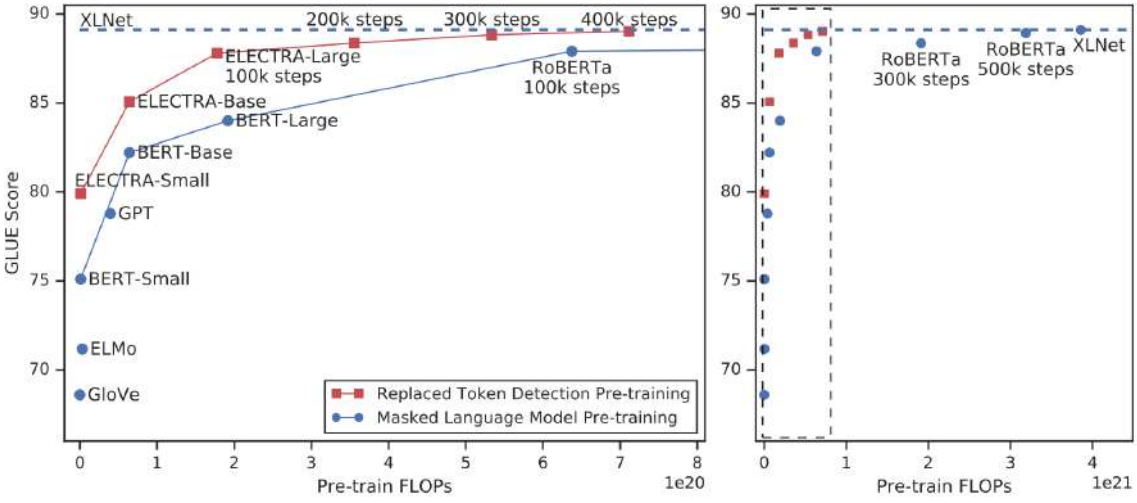


Figura 2.13: Relação entre os custos computacionais (FLOPs) e os *scores* obtidos no *benchmark* GLUE (WANG *et al.*, 2018) para os principais modelos contextuais, evidenciando o expressivo aumento dos custos em relação a incrementos cada vez menores na acurácia, motivando iniciativas de pesquisa por modelos computacionalmente mais eficientes como o ELECTRA.
Fonte: CLARK *et al.* (2020).

Figura 2.14: Tempo de inferência versus tamanho da sentença, com crescimento geométrico em Transformer e crescimento linear na arquitetura Linformer
Fonte: WANG *et al.* (2020).

2.2 Revisão de Literatura

A revisão de literatura foi realizada considerando as principais bases disponíveis no catálogo de periódicos da CAPES[18], em especial as bases Science Direct[19] e Scopus[20] (ambas da *Elsevier*), e Web of Science[21] (*Clarivate Analytics*). Para definição dos critérios de pesquisa, foram considerados os principais termos relacionados às áreas de processamento de linguagem natural, vetorização de palavras, aprendizagem profunda e sua contextualização no domínio de óleo e Gás. Adicionalmente aos termos de busca, a pesquisa foi expandida em função dos autores mais relevantes e com maior número de publicações, identificados a partir das métricas obtidas junto às bases de periódicos. Os artigos foram selecionados em função do impacto de suas contribuições na comunidade científica, por sua aderência ao tema da pesquisa e pela relevância do conteúdo apresentado.

Adicionalmente, a fim de contemplar os aspectos da pesquisa relacionados ao domínio de Óleo e Gás, foram também consideradas as bases de publicações

[18] http://www.periodicos.capes.gov.br/
[19] https://www.sciencedirect.com/
[20] https://www.scopus.com
[21] https://clarivate.com/products/web-of-science/

especializadas como do Instituto Brasileiro de Petróleo[22] (IBP) e a base OnePetro[23] da *Society of Petroleum Engineers* (SPE).

Além das consultas às bases de periódicos anteriormente mencionadas, cabe considerar que a área de processamento de linguagem natural é ricamente coberta por importantes conferências internacionais, onde estão representados o estado-da-arte e as principais contribuições nesse tema. Portanto, a etapa de revisão de literatura concentrou-se muito fortemente em analisar os anais das principais conferências, em especial as organizadas pela *Association for Computational Linguistics* (ACL), que disponibiliza a coletânea dos artigos publicados no repositório *ACL Anthology*[24]. Nesse sentido, as conferências mais relevantes consideradas nesta pesquisa são: *Conference on Computational Natural Language Learning* (CoNLL), *Conference on Empirical Methods in Natural Language Processing* (EMNLP), *Lexical and Computational Semantics and Semantic Evaluation* (SEMEVAL), *North American Chapter of ACL* (NAACL), *Annual Meeting of the Association for Computational Linguistics* (ACL) e *International Conference on Language Resources and Evaluation* (LREC).

Oportunamente, há ainda uma conferência internacional especializada no estudo do tema voltado para o idioma português: a *International Conference on the Computational Processing of Portuguese* (PROPOR). Os anais das edições do evento, disponíveis no repositório Springer[25], também foram analisados.

Em tempo, cabe destacar que a relevância do tema de pesquisa pode ser evidenciada pelo expressivo crescimento no número de publicações nas principais conferências de PLN, conforme reproduzido na Figura 2.15. Da mesma forma, essa tendência é igualmente observada a partir do aumento no número de publicações relacionadas à área de linguística computacional no repositório aberto *arXiv*, mantido pela Cornell University, conforme reportado por PERRAULT *et al.* (2019) e ilustrado na Figura 2.16. De uma maneira geral, toda a área de Inteligência Artificial e seus subdomínios vem apresentando um rápido crescimento em número de participações em conferências (Figura 2.17), sinalizando um crescente interesse nessa área de

[22] https://www.ibp.org.br/publicacoes/
[23] https://www.onepetro.org/
[24] https://aclweb.org/anthology/
[25] https://link.springer.com/conference/propor

conhecimento, conforme destacado pelo relatório anual especializado da Universidade de Stanford (PERRAULT *et al.*, 2019).

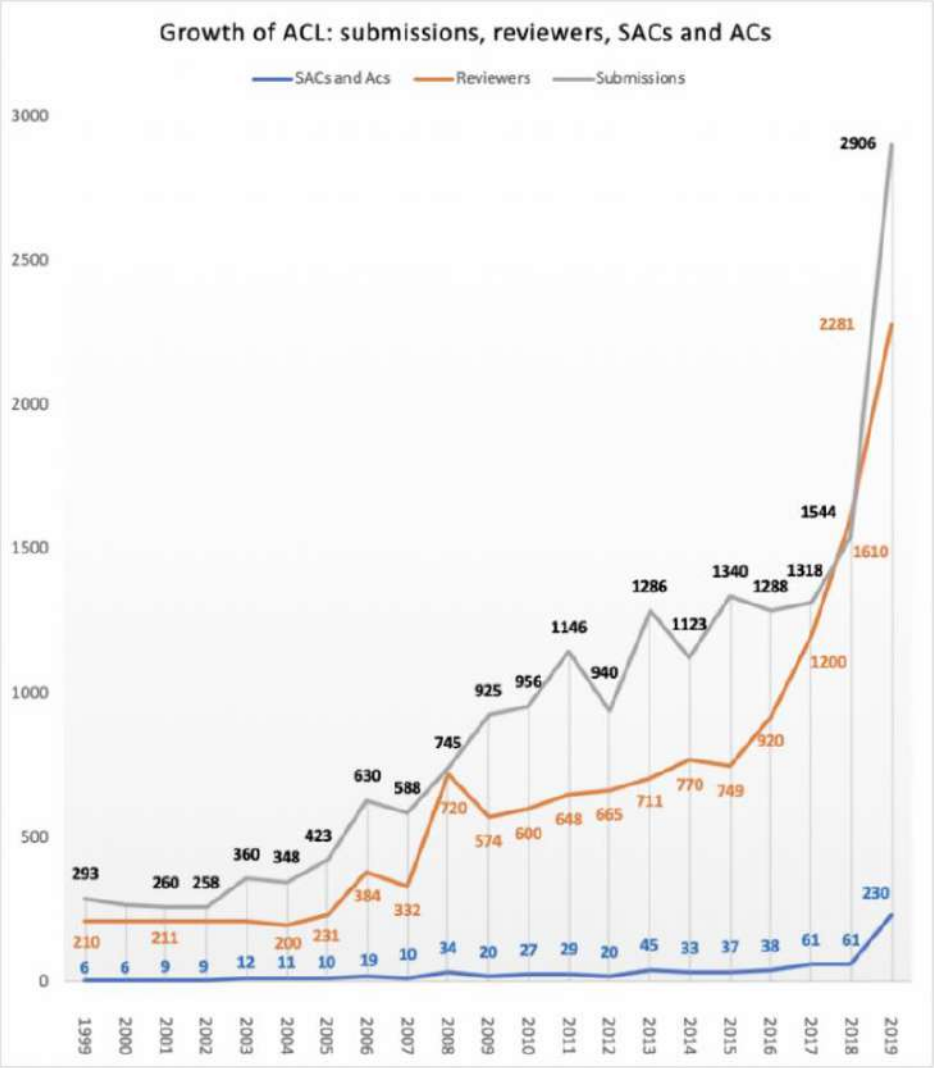


Figura 2.15: Expressivo crescimento de submissões para a conferência da ACL
Fonte: ACL (2019)[26]

[26] https://acl2019pcblog.fileli.unipi.it/?p=156

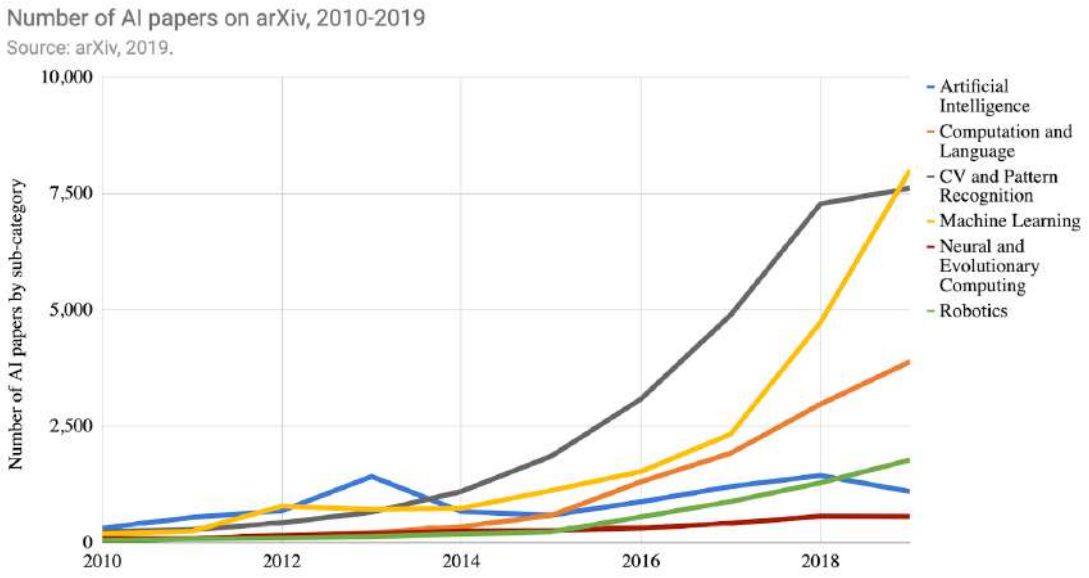


Figura 2.16: Crescimento no número de publicações relacionadas à área de linguística computacional no repositório arXiv
Fonte: PERRAULT *et al.* (2019)

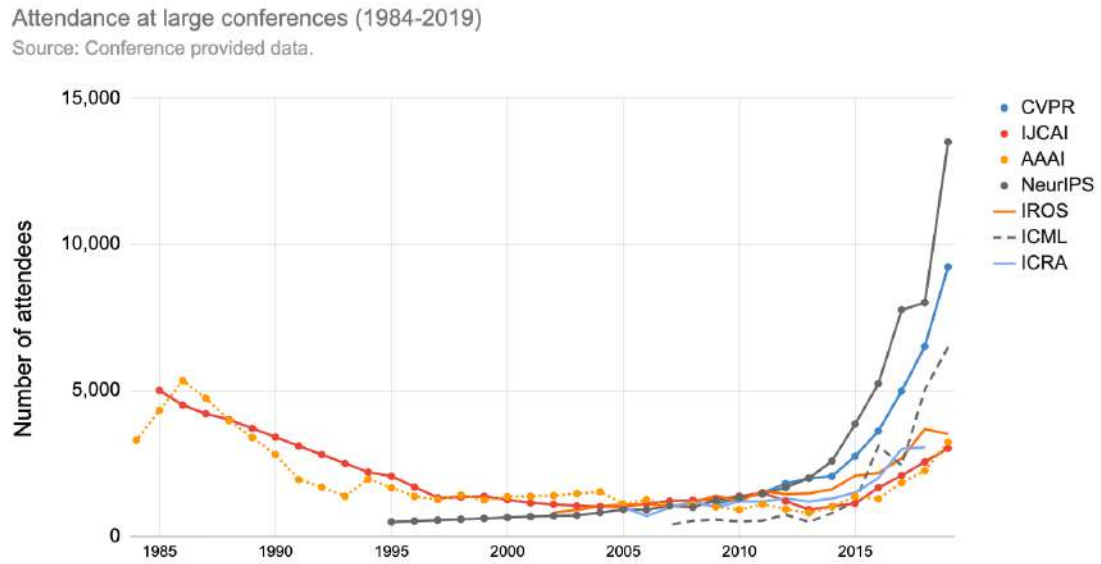


Figura 2.17: Crescimento no número de participantes nas principais conferências de Inteligência Artificial.
Fonte: PERRAULT *et al.* (2019).

Por fim, foram adicionalmente analisadas algumas das principais referências citadas no rico material disponível em alguns dos principais *surveys* identificados no decorrer da pesquisa bibliográfica, como os apresentados por YOUNG *et al.* (2018), KHURANA *et al.* (2018) e CAMACHO-COLLADOS e PILEHVAR (2018), além de livros especializados com publicação recente como JURAVSKY e MARTIN (2020), e PILEHVAR e CAMACHO-COLLADOS (2020).

A Tabela 2.2 relaciona a listagem dos principais artigos identificados e analisados na etapa de revisão de literatura, ordenados em função do ano de publicação.

Tabela 2.2: Principais artigos analisados na revisão de literatura

Ano	Autores	Título
2020	Anna Rogers, Olga Kovaleva, Anna Rumshisky	*A Primer in BERTology: What We Know About How BERT Works*
2020	Simran Arora, Avner May, Jian Zhang, Christopher Ré	*Contextual Embeddings: When Are They Worth It?*
2020	Daniel Jurafsky e James H. Martin	*Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*
2020	Mohammad Pilehvar e Camacho-Collados	*Embeddings in Natural Language Processing: Theory and Advances in Vector Representation of Meaning*
2019	Sebastian Ruder	*Neural Transfer Learning for Natural Language Processing*
2019	Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, C.-C. Jay Kuo	*Evaluating word embedding models: methods and experimental results*
2019	Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, Anubhav Jain	*Unsupervised word embeddings capture latent knowledge from materials science literature*
2019	José Padarian e Ignacio Fuentes	*Word embeddings for application in geosciences development, evaluation, and examples of soil-related concepts*
2019	Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, Donald Brown.	*Text Classification Algorithms: A Survey*
2019	Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova	*Bert: Pre-training of deep bidirectional transformers for language understanding*
2019	Hongfang Lu, Lijun Guo, Mohammad Azim, Kun Huang	*Oil and Gas 4.0 era: A systematic review and outlook*
2018	Farhad Nooralahzadeh, Lilja Øvrelid, Jan Tore Lønning	*Evaluation of Domain-specific Word Embeddings using Knowledge Resources*
2018	Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh	*Natural Language Processing: State of The Art, Current Trends and Challenges*
2018	Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria	*Recent Trends in Deep Learning Based Natural Language Processing*
2018	Jose Camacho-Collados, Mohammad Taher Pilehvar	*From Word to Sense Embeddings: A Survey on Vector Representations of Meaning*
2018	Christina Niklaus, Matthias Cetto, André Freitas, Siegfried Handschuh	*A survey on open information extraction*
2018	Vikas Yadav, Steven Bethard	*A Survey on Recent Advances in Named Entity Recognition from Deep Learning models*
2018	Lei Zhang, Shuai Wang, Bing Liu	*Deep Learning for Sentiment Analysis : A Survey*

2018	Christina Niklaus, Matthias Cetto, André Freitas, Siegfried Handschuh	*A survey on open information extraction*
2018	Sebastian Ruder, Ivan Vulić, Anders Søgaard	*A Survey of Cross-Lingual Word Embedding Models*
2018	João Rodrigues, António Branco	*Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese*
2017	Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jessica S. Rodrigues, Sandra M. Aluísio	*Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks*
2017	Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov	*Enriching Word Vectors with Subword Information*
2017	Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut	*Text Summarization Techniques: A Brief Survey*
2016	Ashwin Ittoo a, Le Minh Nguyen, Antal van den Bosch	*Text analytics in industry: Challenges, desiderata and trends*
2016	Yoav Goldberg	*A primer on neural network models for natural language processing*
2016	Siwei Lai, Kang Liu, Liheng Xu, Jun Zhao	*How to Generate a Good Word Embedding*
2016	Hannah Bast, Björn Buchhold, Elmar Haussmann	*Semantic Search on Text and Knowledge Bases*
2015	Yann LeCun, Y. Bengio, Geoffrey Hinton	*Deep Learning*
2015	Rafal Jozefowicz, Wojciech Zaremba, Ilya Sutskever	*An Empirical Exploration of Recurrent Network Architectures*
2014	Jeffrey Pennington, Richard Socher, Christopher D. Manning	*Glove: Global vectors for word representation*
2013	Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean	*Efficient Estimation of Word Representations in Vector Space*
2013	Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean	*Distributed Representations of Words and Phrases and their Compositionality*
2008	Ronan Collobert, Jason Weston	*A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*

2.2.1 Revisão complementar e relevância do tema

A fim de analisar tendências e estimular novas descobertas em assuntos correlacionados ao objeto de estudo principal da tese, analisamos pesquisas complementares utilizando variações de critérios de busca nas ferramentas Google

Trends[27], Google Acadêmico[28] e Microsoft Academic[29], além dos recursos de *analytics* oferecidos pelas bases Scopus e Web of Science. Essas pesquisas contribuíram para fornecer uma visão mais ampla do espectro de aplicações das técnicas sendo analisadas, inclusive em outros domínios, como nas áreas de biomédica, ciência da computação e engenharias (Figura 2.18 e Figura 2.19).

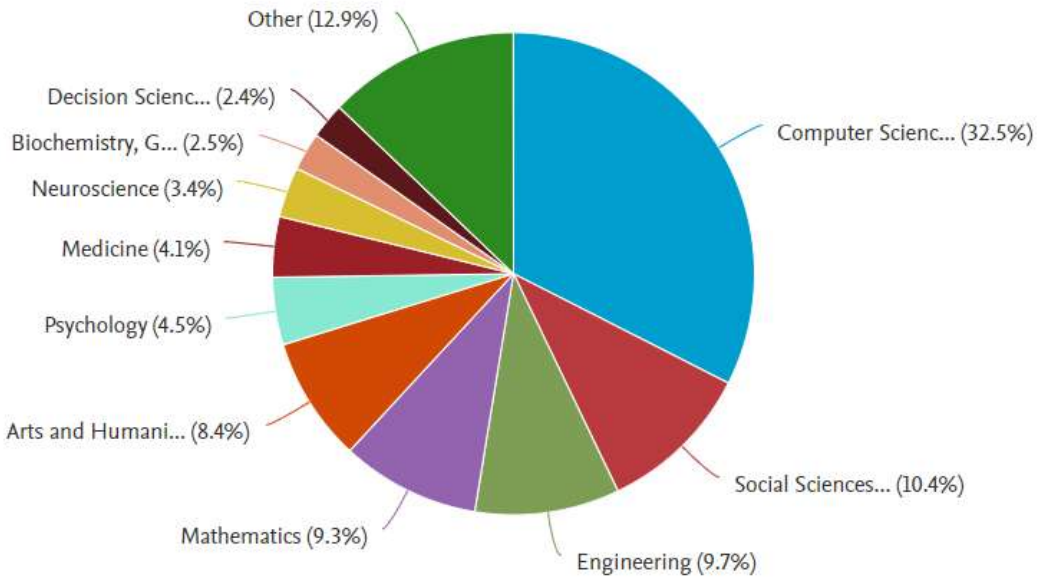


Figura 2.18: Publicações referentes aos termos *"natural language processing"* em diferentes áreas, na base Scopus (janeiro/2021)

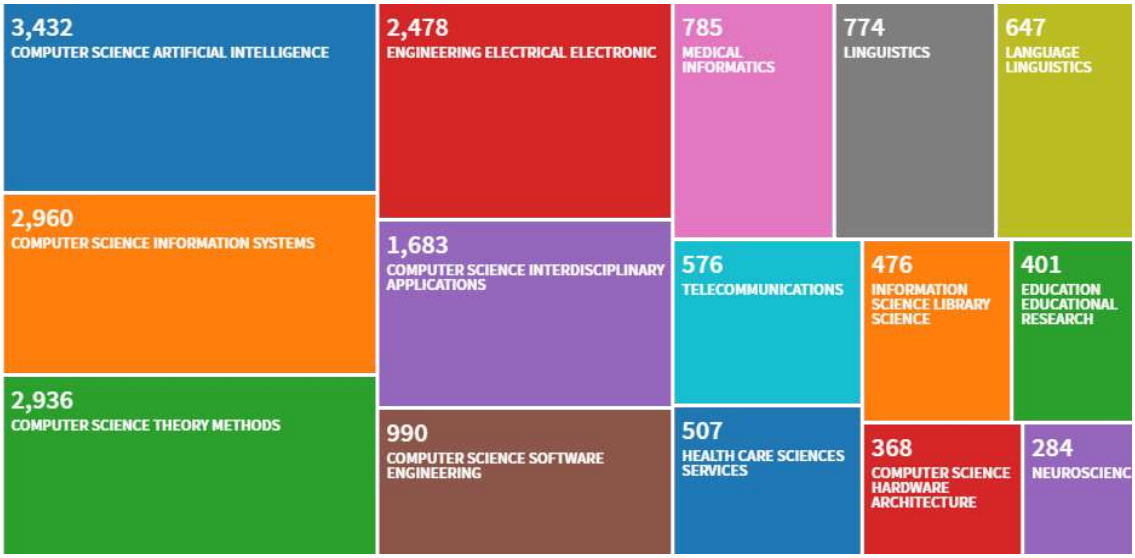


Figura 2.19: Publicações referentes ao tópico *"natural language processing"* em diferentes áreas, na base Web of Science (janeiro/2021)

[27] https://trends.google.com/trends/
[28] https://scholar.google.com.br/
[29] https://academic.microsoft.com/

Primeiramente, com o objetivo de reforçar a relevância do tema de pesquisa da tese, foi analisada a evolução do número de publicações nessa área, utilizando como critério de busca os termos *"deep learning"* e *"natural language processing"*. Observa-se o crescente interesse nessas áreas de conhecimento, evidenciado pelo salto no número de publicações relacionadas a tema, tanto na base Scopus quanto na base Web of Science, conforme ilustrado na Figura 2.20, Figura 2.21 e Figura 2.22.

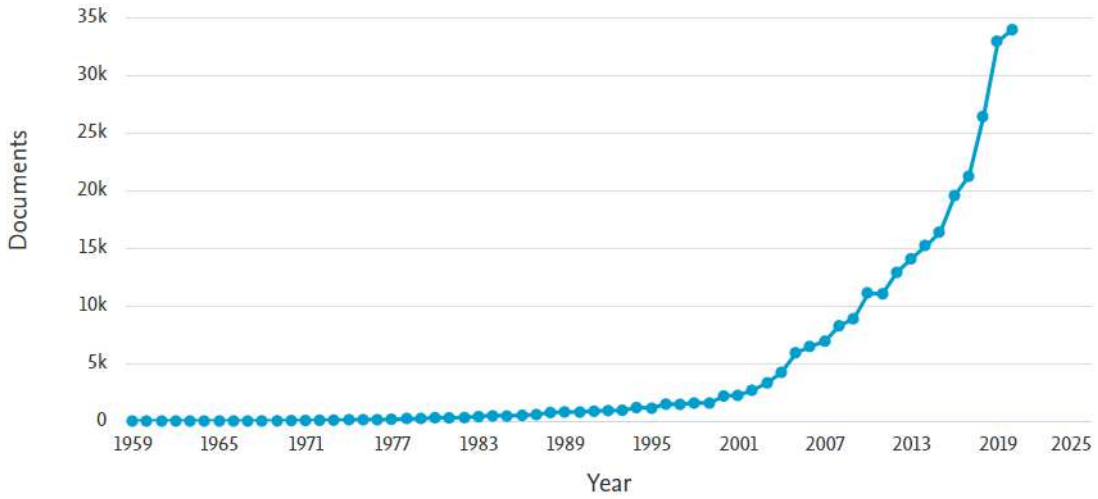


Figura 2.20: Publicações referentes a *"natural language processing"* na base *Scopus*

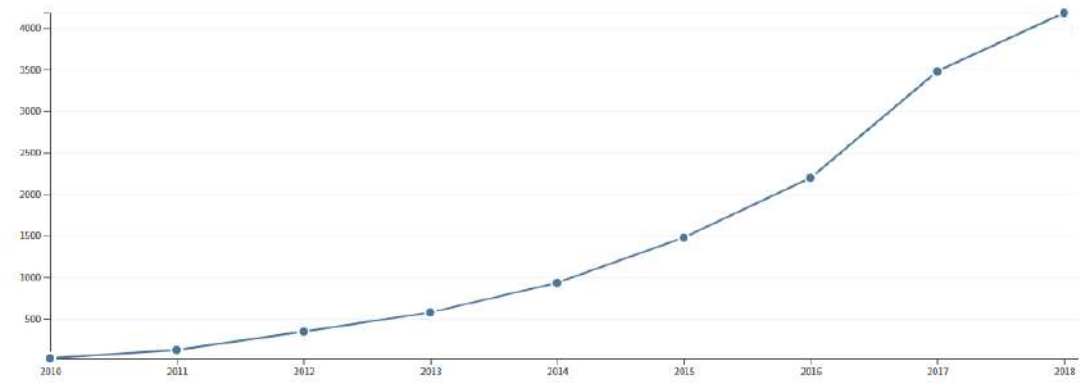


Figura 2.21: Citações por ano sobre o tópico *"natural language processing"* na base *Web of Science*

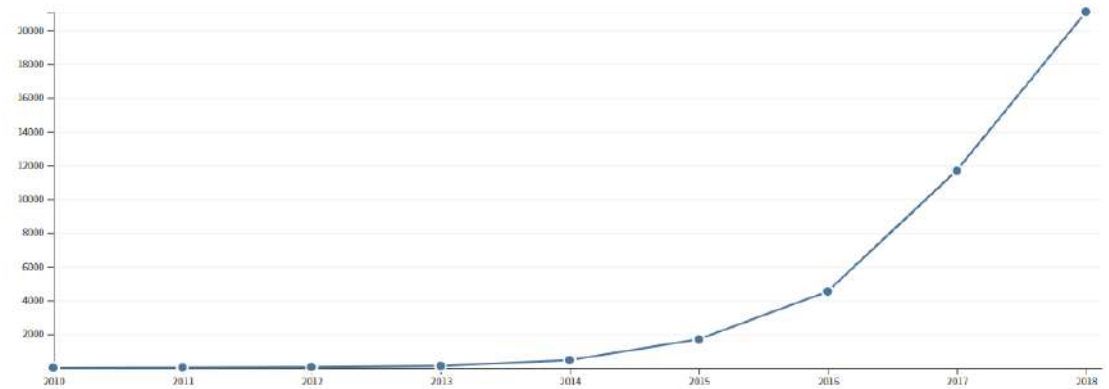


Figura 2.22: Citações por ano sobre o tópico *"deep learning"* na *Web of Science*

Em seguida, foi analisado o cenário atual com as principais publicações nessa área de conhecimento, identificando os *journals*, autores, instituições e conferências mais influentes na área de PLN. Para isso, foi utilizado o Microsoft Academic para gerar relatórios analíticos a partir do tópico *"natural language processing"*, conforme ilustrado na Figura 2.23. De forma complementar, os resultados da pesquisa pelo tópico "*word embeddings*" no Microsoft Academic evidenciam a atual relevância e o crescente interesse neste tema, conforme apresentado na Figura 2.24. Esses resultados auxiliaram no refinamento dos critérios de pesquisa bibliográfica, fornecendo parâmetros mais relevantes para buscar as contribuições de estado-da-arte nessas tecnologias.

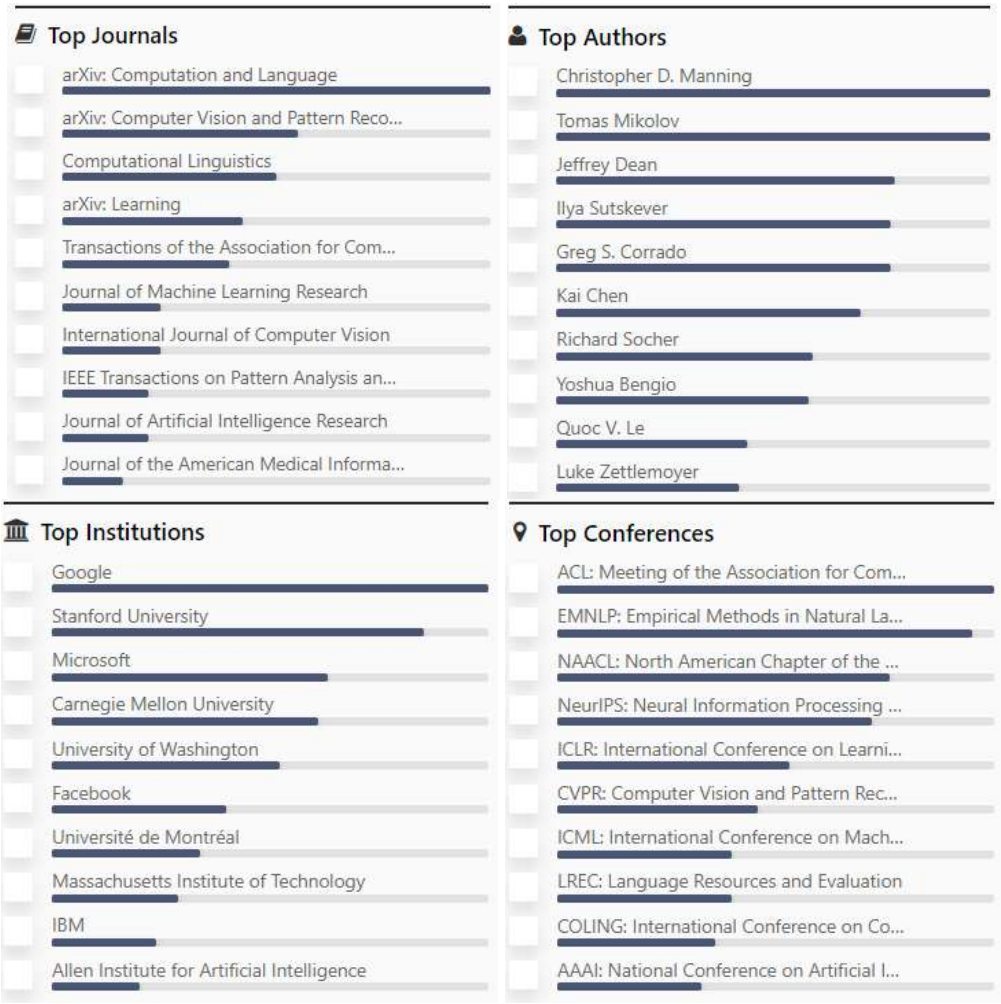


Figura 2.23: Principais *journals*, autores, instituições e conferências mais influentes sobre o tópico *"natural language processing"* no *Microsoft Academic* (janeiro/2021)

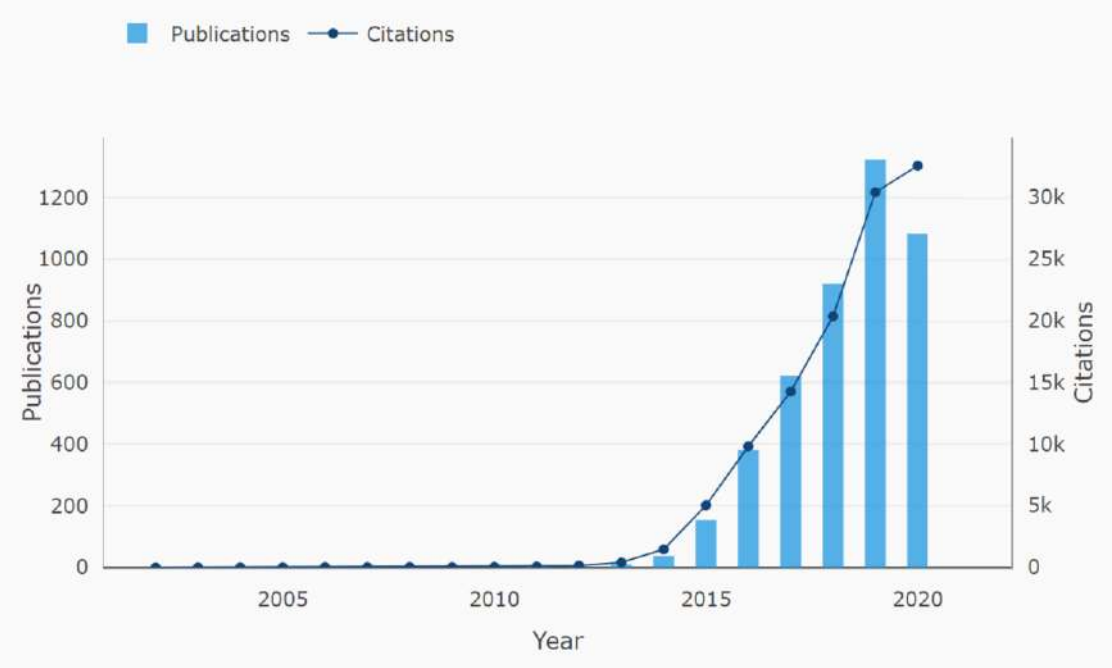


Figura 2.24: Evolução no número de publicações sobre o tópico *'word embeddings'* no *Microsoft Academic* até o ano de 2020 (janeiro/2021)

Em relação ao domínio de Óleo e Gás, uma pesquisa na base OnePetro encontra 4460 artigos associados ao tópico *"natural language processing"* (Figura 2.25), reforçando a relevância dessas técnicas aplicadas ao domínio de O&G. De forma complementar, uma pesquisa nessa base pelo termo '*brazil'* retorna cerca de 10753 publicações, evidenciando a importância da indústria nacional para o cenário global de O&G (Figura 2.26). Em contrapartida, uma pesquisa na base OnePetro pela expressão "*word embeddings*" retorna apenas 12 artigos, um deles escrito em coautoria com o autor desta tese (CORDEIRO *et al.*, 2019), evidenciando que aplicações nessa linha de pesquisa voltadas a situações industriais reais do domínio de O&G ainda são um desafio. Desta forma, após uma abrangente metodologia de pesquisa bibliográfica, e assim ratificando a originalidade e relevância do tema de pesquisa desta tese, não identificamos na literatura científica trabalho similar especializado no domínio de O&G para o idioma português, assim como ficou evidenciada a inexistência de corpora públicos neste domínio, conforme apresentado em detalhes na Seção 2.2.2.

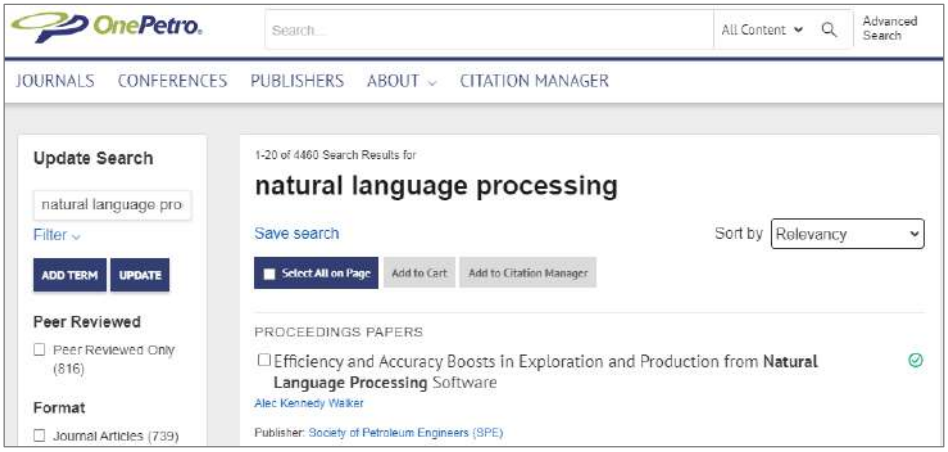


Figura 2.25: Uma consulta na base OnePetro retorna 4460 artigos relacionados ao tópico *"natural language processing"* (janeiro/2021), sugerindo a relevância destas técnicas aplicadas a problemas industriais neste domínio especializado

Figura 2.26: A base OnePetro retorna 10753 artigos relacionados ao termo *"Brazil"* (janeiro/2021), evidenciando a relevância da indústria nacional para o setor de O&G

2.2.2 Trabalhos Relacionados

Em função da grande popularidade recentemente obtida por modelos de vetorização de palavras e suas aplicações em algoritmos de PLN, inúmeros esforços de pesquisa se concentraram em prover modelos pré-treinados a partir de *corpora* de contexto geral. Técnicas de transferência de aprendizado, portanto, se tornaram a opção padrão para aplicações de PLN, viabilizando reutilizar modelos originalmente treinados em domínios genéricos, aplicando-os como dados de entrada para alimentar algoritmos especializados em uma tarefa específica (RUDER *et al.*, 2019).

Entretanto, apesar do sucesso e reconhecimento acadêmico obtidos pelos modelos pré-treinados em domínios genéricos, diversos estudos recentes demonstram que o desenvolvimento de *embeddings* treinados a partir de um corpus especializado no

domínio pode melhorar significativamente a qualidade de suas representações semânticas e, consequentemente, a acurácia das aplicações específicas de PLN em que serão utilizados (LAI *et al.*, 2016; PAKHOMOV *et al.*, 2016; NOORALAHZADEH *et al.*, 2018; WANG *et al.*, 2018; ALSENTZER *et al.*, 2019; TSHITOYAN *et al.*, 2019; PADARIAN e FUENTES, 2019; GOMES *et al.*, 2021).

Portanto, o desenvolvimento e avaliação de modelos vetoriais especializados em domínio específico representam uma importante área de pesquisa em PLN. LAI *et al.* (2016) apresentaram um detalhado estudo sobre as melhores práticas para garantir a qualidade na geração de *embeddings* especializados, destacando que "*o domínio do corpus é mais importante que o seu tamanho*" e recomendam "*escolher o corpus em um domínio adequado para a tarefa desejada*". Esses modelos especializados têm sido amplamente utilizados na área acadêmica e em diversas aplicações industriais. KHABIRI *et al.* (2019) aplicaram *embeddings* especializados em problemas de classificação de *logs* na indústria, e MISRHA e SHARMA (2019) os utilizaram para identificar com sucesso ambiguidades em requisitos de *software*. TSHITOYAN *et al.* (2019) apresentaram impressionantes resultados sobre a utilização de *word embeddings* treinados a partir de literatura de ciência de materiais, capazes de predizer potenciais descobertas sobre novas propriedades termoelétricas de materiais com anos de antecedência, e reportaram que "*a qualidade e a especificidade do corpus no domínio são fatores determinantes para a utilidade dos embeddings em tarefas especializadas*".

O domínio biomédico, em particular, face à ampla disponibilidade de corpora e conjuntos de dados anotados para avaliação, representa um dos domínios especializados mais ativos em pesquisas na área de PLN (JIANG *et al.*, 2015; WANG *et al.*, 2018; ALSENTZER *et al.*,2019; LEE *et al.*, 2020). KALYAN e SANGEETHA (2020) apresentaram uma extensa revisão sobre as aplicações de *word embeddings* para PNL na área médica.

Com relação ao domínio de O&G, apenas recentemente alguns poucos estudos se propuseram a desenvolver modelos especializados, majoritariamente no idioma inglês, principalmente em função da característica escassez de dados de treinamento publicamente disponíveis. NOORALAHZADEH *et al.* (2018) apresentam um estudo sobre a geração e avaliação de modelos de *embeddings* para O&G no idioma inglês. O estudo apresenta resultados para avaliações intrínsecas considerando um conjunto de

testes construído a partir do *Schlumberger Oilfield Glossary*[30], medindo acurácia, precisão e revocação dos modelos para termos semanticamente relacionados. Os autores reportam resultados superiores em comparação a um modelo de contexto genérico, e concluem que a construção de modelos especializados é vantajosa mesmo considerando a limitada disponibilidade de corpora especializados. PADARIAN e FUENTES (2019) apresentaram *word embeddings* em inglês para o domínio de Geociências, assim como uma suíte de testes para avaliação intrínseca. Os modelos especializados foram comparados a um modelo de domínio genérico, obtendo avanços significativos em tarefas específicas como analogias semânticas e categorização.

Todos os estudos previamente mencionados reiteram que, para domínios específicos como o de O&G, o desenvolvimento de modelos de vetorização de palavras especializados justifica-se mesmo para os casos em que os dados disponíveis são consideravelmente menores do que os corpora de contexto genérico. Especificamente para o idioma português, uma iniciativa pioneira foi apresentada pelo autor desta tese (GOMES *et al.*, 2018), desenvolvendo o único conjunto de *word embeddings* em português para o domínio O&G, treinados em uma versão significativamente menor do corpus atual. Esses modelos foram aplicados em um trabalho de reconhecimento de entidades nomeadas, com um artigo publicado em coautoria na conferência LREC (CONSOLI *et al.*, 2020). Adicionalmente, os modelos vetoriais também estão implantados no ambiente de produção da Petrobras para implementar funcionalidades de expansão automática de consulta em problemas de busca semântica, conforme publicação em coautoria com KRUEL *et al.* (2019). Uma abordagem relacionada para vetorização de documentos foi utilizada em experimentos na área de inteligência tecnológica, apresentado em coautoria em uma publicação na conferência *Offshore Technology Conference* (CORDEIRO *et al.*, 2019). Nesta tese, portanto, objetiva-se evoluir a partir dos resultados preliminarmente apresentados, expandindo significativamente o corpus especializado de treinamento, além de oferecer dados de testes e metodologias para avaliação intrínseca e extrínseca, novas análises qualitativas e explorando uma análise comparativa com um modelo de referência treinado em corpora de contexto geral.

[30] Schlumberger Oilfield Glossary. Disponível em: https://www.glossary.oilfield.slb.com/

Capítulo III

Corpora de O&G e Modelos Vetoriais

"Data is the new Oil. It´s valuable, but if unrefined it cannot really be used."

Clive Humby

Este capítulo descreve a metodologia para a coleta, organização e tratamento do conjunto de dados textual utilizado para compor o corpus especializado de Óleo e Gás em português. Em seguida, descrevemos o processo de treinamento e os parâmetros utilizados na geração dos modelos vetoriais **`PetroVec`**.

3.1 Visão Geral

Modelos de vetorização de palavras buscam atribuir uma representação vetorial para cada palavra de um vocabulário, capturando relações de similaridade sintática e semântica a partir do contexto em que ocorrem em um corpus. Essas técnicas consistem em representar cada termo como um vetor contínuo *n*-dimensional de valores reais, de maneira que palavras relacionadas entre si sejam posicionadas em regiões próximas no espaço vetorial criado.

O treinamento desses modelos vetoriais demanda um grande volume de dados no formato texto, preferencialmente representativo do domínio em que serão aplicados. Porém, muitas vezes esses dados não estão disponíveis em quantidade e qualidade suficientes, especialmente em um determinado domínio e/ou idioma. Uma técnica comumente aplicável é denominada transferência de aprendizado, que consiste em reutilizar nos algoritmos de PLN modelos vetoriais pré-treinados a partir de um outro corpus de domínio geral (CER *et al.*, 2018). Esses modelos públicos de contexto geral são comumente denominados *genéricos* ou *globais*.

Entretanto, estudos sugerem que treinar os modelos a partir de um corpus específico do domínio pode melhorar significativamente a qualidade semântica das representações vetoriais e, consequentemente, o desempenho dos algoritmos de PLN em que serão aplicados (LAI *et al.*, 2016; NOORALAHZADEH *et al.*, 2018; TSHITOYAN *et al.*, 2019; PADARIAN e FUENTES, 2019). Nesse cenário, os vetores podem ser treinados a partir de elementos textuais mais representativos do domínio e, portanto, conseguem representar mais adequadamente os termos técnicos específicos de seu vocabulário. Em um determinado contexto técnico, uma palavra pode assumir um significado completamente distinto, demandando assim representações adequadas a essa especialidade. Esses modelos treinados a partir de corpora de domínio são chamados de *especializados* ou *locais*.

Portanto, este capítulo se propõe a descrever o processo de criação do corpus especializado e a geração do **`PetroVec`**, um conjunto de modelos de vetorização de palavras especializado no domínio de O&G para o idioma português. Os modelos são gerados utilizando dois dos principais algoritmos disponíveis: Word2vec (MIKOLOV *et al.*, 2013a) e FastText (BOJANOWSKI *et al.*, 2017), considerando diferentes composições de corpora para analisar sua influência na qualidade dos modelos.

3.2 O *Corpus* de O&G em português

Considerando a escassez de corpora públicos em português para o domínio técnico de O&G na literatura científica, um aspecto fundamental do escopo deste trabalho consiste em coletar, organizar e processar uma grande coleção textual especializada neste domínio. Essa coleção objetiva constituir um corpus especializado de referência neste domínio em português, composto por milhares de documentos técnicos e científicos originalmente publicados pelas principais Universidades e demais instituições de referência atuantes área de petróleo no idioma português, notadamente a Petrobras, a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

O corpus inicial foi formado a partir de um conjunto de 202 boletins técnicos publicados pela Petrobras[31], contendo ao todo cerca de 2000 artigos na área de Petróleo. Em seguida, incorporamos um conjunto de 727 documentos publicados pela ANP, correspondendo a 420 documentos acadêmicos (teses, dissertações e monografias na área de petróleo, elaboradas no contexto de seu Programa de Recursos Humanos-PRH[32]), e mais 307 publicações periódicas, notas e estudos técnicos[33]. Para enriquecimento do vocabulário técnico, foi incluído um glossário da ANP[34] (contendo aproximadamente 560 termos técnicos), além de dois outros recursos obtidos junto ao Dicionário do Petróleo (FERNÁNDEZ *et al.*, 2009), sendo um glossário[35] (com 8860 termos) e um dicionário de siglas[36] (contendo 1257 termos). Adicionalmente, outras bases foram adicionadas ao corpus, incluindo artigos técnicos disponíveis nos anais do Congresso da Rio Oil and Gas[37], promovido pelo Instituto Brasileiro de Petróleo (IBP). Um conjunto de 2.067 documentos acadêmicos foi obtido junto à Memória Técnica da Petrobras, composto por teses e dissertações de empregados da Companhia com temas relacionados à área de petróleo. Por fim, o corpus foi significativamente enriquecido com a inclusão de um grande conjunto de documentos acadêmicos fornecidos por CORDEIRO (2020), contemplando cerca de 3200 teses e dissertações coletadas a partir da Biblioteca Digital Brasileira de Teses e Dissertações[38], do Instituto Brasileiro de Informação em Ciência e

[31] http://publicacoes.petrobras.com.br
[32] http://www.anp.gov.br/pesquisa-desenvolvimento-e-inovacao/prh-anp-programa-de-formacao-de-recursos-humanos
[33] http://www.anp.gov.br/notas-tecnicas
[34] http://www.anp.gov.br/glossario
[35] http://dicionariodopetroleo.com.br/glossario/
[36] http://dicionariodopetroleo.com.br/siglario/
[37] https://www.ibp.org.br/rog2018-trabalhos-tecnicos/
[38] http://bdtd.ibict.br/

Tecnologia, selecionados por algoritmos automáticos de classificação conforme relevância e aderência ao domínio de O&G. Dessa forma, compilamos um corpus representativo do domínio de O&G em português, a partir de agora referenciado como *corpus especializado*, composto por artigos científicos, teses, dissertações, glossários, periódicos, relatórios e notas técnicas, contemplando cerca de 85.7 milhões de *tokens (*um *token* pode ser entendido como a unidade mínima de representação em PLN, que neste trabalho representa cada ocorrência de uma palavra do conjunto textual).

A fim de evitar quaisquer padrões linguísticos incorretos, indevidamente introduzidos por algoritmos de tradução automática, nossos corpora são compostos apenas por textos originalmente escritos em português e publicados pelas instituições anteriormente mencionadas, que operam nativamente neste idioma no domínio de O&G. Portanto, nenhum processo de tradução automática foi realizado no material utilizado neste trabalho.

Adicionalmente, para ampliar a cobertura dos aspectos linguísticos do idioma e melhorar as representações semânticas inclusive para o vocabulário não-técnico, incluímos um corpus de contexto geral de referência em português fornecido por HARTMANN *et al.* (2017). A inclusão desse corpus genérico teve como objetivo permitir a análise da qualidade de modelos híbridos (ou seja, contendo tanto o corpus especializado em O&G como o corpus de contexto geral), e a influência de variações de corpora na qualidade dos modelos. Especificamente sobre este corpus (a partir de agora referenciado como *corpus genérico*), obtivemos apenas uma fração de cerca de 50% do conteúdo originalmente reportado no trabalho de HARTMANN *et al.* (2017), correspondendo à parte pública disponibilizada pelos autores.

A Tabela 3.1 descreve informações detalhadas sobre a composição dos corpora, suas correspondentes fontes de coleta, além da contabilização total do número de sentenças e de tokens após o pré-processamento dos textos. A Figura 3.1 ilustra a proporção de cada base na composição total do *corpus especializado*. Até o presente momento, o *corpus especializado* é o maior conjunto textual já reportado para o domínio específico de O&G em português. Objetiva-se que o compartilhamento público deste material com a comunidade científica possa contribuir com o avanço da linha de pesquisa de PLN especializada neste domínio.

Tabela 3.1: Composição dos *corpora* adquiridos para este trabalho

Domínio	Fonte	Descrição	Sentenças	Tokens
Especializado	Petrobras	Boletins Petrobras de Geociências e de Produção de Petróleo	298.865	3.821.966
		Teses e dissertações no domínio de O&G	2.939.262	37.027.438
	ANP	Boletins e relatórios técnicos	132.955	2.136.465
		Teses e dissertações no domínio de O&G	279.196	3.629.999
	IBP	Anais da conferência Rio O&G	89.116	1.287.223
	IBICT-BDTD	Teses e dissertações filtradas por assuntos relacionados ao domínio de O&G	2.558.837	37.825.743
	Total		***6.298.231***	***85.728.834***
Genérico	HARTMANN *et al.* (2017)	Fração pública do corpus de contexto geral em português	37.327.741	365.295.169
Híbrido		Combinação dos corpora Genérico e Específico	43.622.972	451.021.003

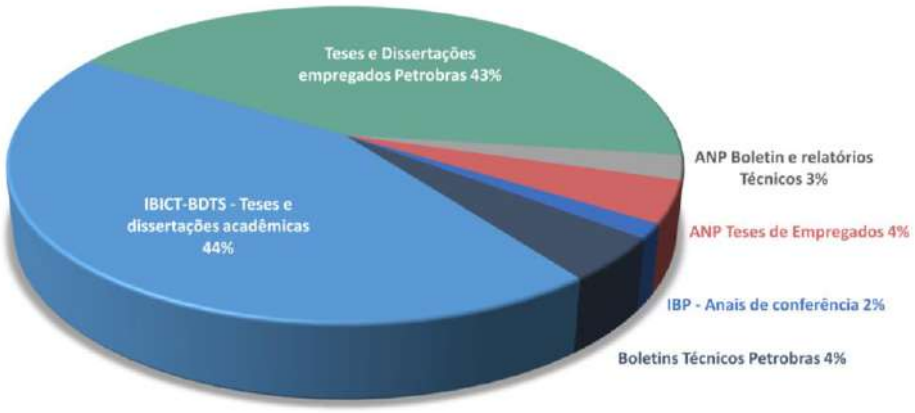


Figura 3.1: Proporção do tamanho de cada base (em número de tokens) na composição do *corpus especializado*

Em complemento ao conjunto textual composto exclusivamente a partir de fontes públicas, passíveis de compartilhamento com a comunidade científica, no decorrer do processo de coleta para viabilizar o objeto de estudo desta tese uma significativa coleção complementar de documentos foi obtida diretamente junto à Petrobras. Trata-se de um rico conteúdo representativo do domínio, composto por milhares de documentos armazenados na Memória Técnica da Petrobras (como artigos, relatórios técnicos e material científico produzido internamente), na Base de Dados da Exploração e Produção (E&P), e no repositório de publicações Petrosin, conforme descrito na Tabela 3.2. Essa inclusão resultou em um expressivo aumento na representatividade do *corpus especializado*, acrescentando mais de 273 milhões de *tokens*. No entanto, esses dados não estão disponíveis para o público externo, sendo classificados como restritos para uso em iniciativas conduzidas internamente na Petrobras. Portanto, devido às limitações para divulgação e compartilhamento de resultados junto à comunidade científica, esse *corpus*

restrito não está contemplado no escopo desta tese, sendo abordado como possíveis trabalhos futuros para a continuidade do treinamento, contemplando essa rica fonte especializada. A Figura 3.2 ilustra uma simulação de como seria a composição proporcional total do corpus especializado, considerando uma eventual inclusão das bases restritas.

Tabela 3.2: Corpora restritos, para uso interno na Petrobras

Fonte	Descrição	Sentenças	Tokens
Petrobras (uso restrito)	Memória Técnica – Comunicados	1.600.963	21.192.511
	Memória Técnica – Relatórios técnicos	3.088.305	39.895.527
	Base Integrada de Arquivos de E&P	21.422.868	178.564.938
	Petrosin – Artigos de empregados Petrobras	2.735.396	33.912.420
Total		***28.847.532***	***273.565.396***

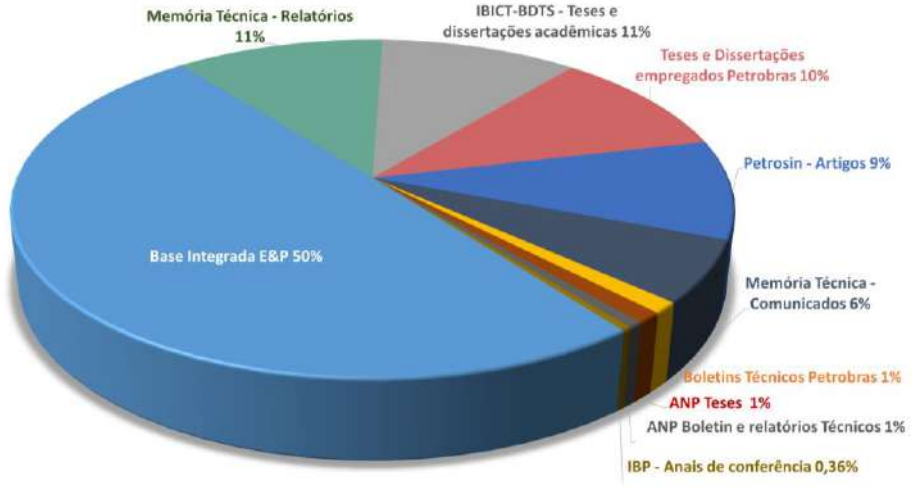


Figura 3.2: Proporção do tamanho de cada base (em número de tokens), simulando a inclusão das bases restritas (em cinza) na composição do *corpus especializado*

3.3 Extração e pré-tratamento dos dados

Na fase de coleta de dados, os documentos foram originalmente obtidos em seus formatos nativos, em sua maioria PDF e Microsoft Word. Portanto, os arquivos foram posteriormente convertidos para o formato texto simples, através de um algoritmo de extração desenvolvido em Java com a ferramenta Apache Tika (MATTMANN e ZITTING, 2011). A realização de experimentações adicionais sobre melhorias de extração com técnicas de reconhecimento de caracteres (*Optical Character Recognition*, *OCR*) estão fora do escopo deste estudo, embora possam futuramente contribuir para uma melhor qualidade do texto extraído.

Em seguida, o corpus passou pela fase de pré-tratamento, para eliminação de ruídos, normalização léxica, segmentação de sentenças e preparação dos dados de uma

forma adequada para treinamento pelos algoritmos de aprendizagem automática, considerando as melhores práticas para o idioma português descritas por HARTMANN *et al.* (2017) e RODRIGUES e BRANCO (2016). Inicialmente, utilizamos diretamente os *scripts* fornecidos por HARTMANN *et al.* (2017), disponíveis no repositório Github do artigo[39]. No entanto, em função da diferente natureza e formato dos arquivos texto utilizados por aquele projeto, composto notadamente por conteúdo obtido de páginas da internet, os *scripts* se mostraram inadequados ao nosso formato de textos. Portanto, um novo código foi desenvolvido especificamente para atender aos propósitos deste trabalho, considerando, ainda assim, as principais recomendações propostas por esses autores.

Nesse contexto, diversas técnicas para normalização léxica e limpeza dos dados foram aplicadas. Primeiramente, todos os caracteres foram convertidos para sua representação minúscula, e foram eliminados caracteres especiais e de pontuação. Observou-se a princípio uma ampla divergência no formato de palavras acentuadas, o que motivou a uniformização dessa grafia substituindo os diacríticos por sua forma equivalente não-acentuada. Palavras muito comuns (*stopwords*), que para os propósitos deste trabalho não contribuem com informação útil, também foram descartadas utilizando a biblioteca Python NLTK (LOPER e BIRD, 2002). Tokens exclusivamente compostos por caracteres numéricos foram substituídos pela *tag* <NUMBER>, e múltiplas ocorrências consecutivas dessa mesma *tag* foram substituídas por apenas uma. Finalmente, o texto pré-processado e limpo foi segmentado em sentenças (etapa de *tokenização*) e linhas contendo menos de três palavras foram descartadas.

No decorrer do tratamento do *corpus*, cabe ressaltar a observação da ocorrência de muitas palavras raras (com frequência igual a 1), potencialmente associadas a erros de grafia resultantes de problemas da extração textual a partir dos arquivos PDF. Esses termos foram desconsiderados durante a etapa de treinamento, porém o problema evidencia a necessidade de melhores técnicas de extração para os documentos, buscando aumentar a qualidade final do *corpus*, a ser endereçada em trabalhos futuros.

3.4 Treinamento dos modelos PetroVec

Os modelos **`PetroVec`** foram treinados utilizando os algoritmos Word2vec (MIKOLOV et al., 2013a) e FastText (BOJANOWSKI et al., 2017). Optamos por utilizar

[39] https://github.com/nathanshartmann/portuguese_word_embeddings

as principais definições de hiperparâmetros conforme previamente recomendado em outros trabalhos (MIKOLOV *et al.*, 2013; LAI *et al.*, 2016; HARTMANN *et al.*, 2017; NOORALAHZADEH *et al.*, 2018), considerando que o objetivo principal deste trabalho é fornecer um conjunto de modelos vetoriais pré-treinados para domínio específico, em vez de apresentar novidades técnicas sobre métodos e algoritmos de treinamento. Isto é, utilizamos tamanho da janela = 5, método *skip-gram, e* vetores com 100 dimensões que, segundo LAI *et al*. (2016), podem oferecer um balanço adequando entre a qualidade dos modelos e o custo computacional, sendo suficientes para fornecer desempenho satisfatório para a maior parte das aplicações de PLN. Palavras que ocorrem menos de 10 vezes no corpus também foram descartadas durante o treinamento.

A fim de viabilizar uma avaliação comparativa dos efeitos de diferentes corpora na qualidade dos modelos resultantes, geramos versões dos modelos considerando duas variações de composição de corpora, conforme anteriormente apresentados na Tabela 3.1 e identificados como *Corpus Especializado* e *Corpus Híbrido*. Isto é, para cada algoritmo (Word2vec e FastText), foram gerados dois modelos: um contemplando apenas o corpus especializado de O&G, e outra contemplando o conjunto total de documentos, inclusive o corpus genérico. Portanto, o modelo treinado a partir do *corpus específico* será referenciado como **`PetroVec-O&G`**, enquanto o modelo gerado a partir do *corpus híbrido* será referenciado como **`PetroVec-hybrid`**.

Adicionalmente, para estabelecer um *baseline* comparativo para as métricas de avaliação dos modelos **`PetroVec`**, utilizamos como referência um modelo pré-treinado de contexto geral em português fornecido por HARTMANN *et al.* (2017), obtido a partir do repositório de *word embeddings* disponibilizado pelo Núcleo Interinstitucional de Linguística Computacional (NILC) [40] . Esse modelo genérico, a partir de agora referenciado como **`skipgram-NILC`**, foi treinado a partir de uma grande coleção textual de domínio geral, contendo cerca de 1,4 bilhões de tokens.

Por fim, para oferecer um conjunto mais representativo de avaliações comparativas dos modelos **`PetroVec`**, também foi utilizado um modelo de *embedding* para O&G apresentado em um trabalho preliminar do autor desta tese (GOMES *et al.*, 2018), treinado em uma versão significativamente menor do corpus especializado. Nesta tese, esse modelo preliminar será referenciado como **`baseline-O&G`**.

[40] http://nilc.icmc.usp.br/embeddings

A Tabela 3.3 apresenta informações detalhadas sobre cada um dos modelos utilizados nesta tese, incluindo suas respectivas composições de corpora, contagem de tokens e tamanho do vocabulário contabilizados após as etapas de pré-processamento descritas na Seção 3.3. Uma representação gráfica comparativa desses dados é apresentada na Figura 3.3.

Tabela 3.3: Composição de *corpora* para cada modelo referenciado neste trabalho

Modelo	Descrição	Tamanho vocabulário	Tamanho do corpus
skipgram-NILC	Modelo público, pré-treinado a partir de corpora genérico para o idioma português (HARTMANN *et al.*, 2017)	929.606	1.395.926.282
baseline-O&G	Modelo especializado para O&G utilizado como baseline, apresentado em um trabalho preliminar deste autor (GOMES *et al.*, 2018) e treinado a partir de uma versão significativamente menor do corpus específico.	113.934	10.109.732
PetroVec-O&G	Modelo treinado no *Corpus Especializado* de O&G	161.842	85.725.834
PetroVec-Hybrid	Modelo treinado a partir do *Corpus Híbrido*, contemplando tanto o corpus especializado como o corpus genérico.	440.692	451.021.003

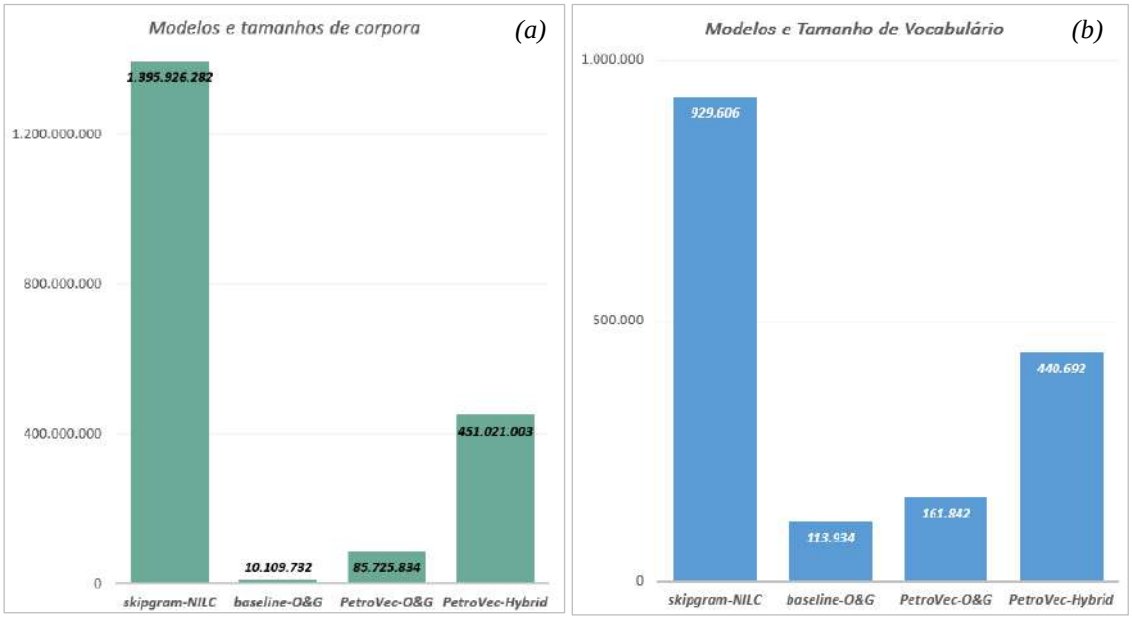


Figura 3.3: Gráfico comparativo com as diferentes dimensões dos corpora utilizados em cada modelo, em contagem de tokens *(a)* e tamanho do vocabulário *(b)*

3.4.1 Ambiente Computacional

Os experimentos foram majoritariamente realizados em equipamento com processador i7, 32GB de RAM e placa GPU Nvidia Quadro M2000 com 4GB, sistema operacional Windows 10.

A plataforma Java 8 foi utilizada para implementação dos scripts de extração do conteúdo textual dos arquivos nos formatos PDF e Word, utilizando a ferramenta Apache Tika[41] (MATTMANN e ZITTING, 2011). A plataforma Python foi usada nas demais etapas de pré-processamento, treinamento e avaliação dos resultados dos experimentos. Para a *tokenização*, remoção de *stopwords* e normalização de caracteres acentuados, utilizou-se a biblioteca NLTK[42] (LOPER e BIRD, 2002). O treinamento dos modelos vetoriais foi realizado usando a biblioteca Gensim[43] (REHUREK e SOJKA, 2011), disponível gratuitamente para a linguagem Python e referência no ambiente acadêmico. O motivo para a escolha dessa ferramenta é sua versatilidade ao permitir o treinamento e a leitura dos vetores de maneira compatível com diferentes algoritmos.

[41] https://tika.apache.org/
[42] https://www.nltk.org/
[43] https://radimrehurek.com/gensim/

Capítulo IV

Avaliação dos Modelos PetroVec

> "Não se imagina como tudo é vago, até que se tente fazê-lo de maneira precisa."
>
> Bertrand Russell

Este capítulo apresenta o conjunto de diferentes metodologias de avaliações desenvolvidas para o **`PetroVec`**. Os modelos são avaliados segundo métricas quantitativas (baseados em análises intrínsecas e extrínsecas), além de oferecer um amplo conjunto de análises qualitativas, para avaliar propriedades linguísticas codificadas no espaço semântico dos modelos. Os resultados obtidos são apresentados e discutidos, oferecendo uma análise comparativa em relação a um modelo público de contexto geral de referência para o idioma português, além de um modelo especializado para O&G resultado de um trabalho preliminar, para servirem como base de referência para as métricas obtidas.

4.1 Visão Geral

Apesar da crescente popularidade obtida por modelos de vetorização de palavras e de inúmeras aplicações em algoritmos de PLN reportadas com sucesso em recentes publicações, ainda não há um consenso na comunidade científica sobre um método único a ser utilizado para prover uma adequada avaliação da qualidade dessas representações semânticas de linguagem (SCHNABEL *et al.*, 2015; BARAKOV, 2018), uma vez que cada abordagem de avaliação foca em um aspecto linguístico diferente dos modelos (WANG *et al.*, 2019), principalmente quando aplicados a um domínio específico (NOORALAHZADEH, 2020). Há, entretanto, uma variedade de métodos e *benchmarks* utilizados para avaliações dos modelos[44] (ROGERS *et al.*, 2019), predominantemente voltados de domínios genéricos, cujas abordagens amplamente aceitas podem ser classificadas em duas categorias principais: avaliações *intrínsecas* e *extrínsecas* (SCHNABEL *et al.*, 2015).

A avaliação *intrínseca* (apresentada na Seção 4.2) objetiva estabelecer uma métrica para medir a qualidade das representações semânticas internas dos modelos, correlacionando a similaridade obtida pelos modelos em relação a um *dataset* de referência manualmente anotado. A avaliação *extrínseca* (detalhada na Seção 4.3), por sua vez, consiste em medir a contribuição de um modelo de vetorização quando aplicado a uma tarefa específica de PLN (como classificação de texto, REN, tradução automática etc.), avaliando a melhora ou piora do algoritmo no desempenho da tarefa.

Neste capítulo, apresentamos uma extensa cobertura de avaliações para os modelos **`PetroVec`**, contemplando metodologias quantitativas (baseadas em análise intrínseca e extrínseca), além de oferecer uma variedade de análises qualitativas, para avaliar aspectos de linguagem codificados no espaço semântico dos modelos. Os resultados obtidos são analisados e discutidos, fornecendo uma análise comparativa em relação a um modelo público de referência, pré-treinado em corpora de domínio genérico em português (HARTMANN *et al.*, 2017). Adicionalmente, os resultados são comparados a um modelo especializado de O&G desenvolvido em um trabalho preliminar do autor desta tese (GOMES *et al.*, 2018), para servir como *baseline*.

[44] Há um evento científico especializado sobre esse tema: *RepEval: Workshop on Evaluating Vector Space Representations for NLP*. Disponível em: https://www.aclweb.org/anthology/venues/repeval/

Os resultados das análises convergem ao evidenciar que os modelos especializados são capazes de capturar nuances de representação de uma forma muito mais fiel às características específicas do domínio, mantendo-se vantajosos mesmo diante de uma menor disponibilidade de corpora para treinamento. As seções a seguir apresentam as metodologias de avaliação e seus respectivos resultados.

4.2 Avaliação intrínseca

A avaliação *intrínseca* objetiva atribuir uma métrica sobre a qualidade dos modelos de *embedding* e sua capacidade de capturar propriedades sintáticas e semânticas a partir do corpus, medindo as relações de similaridade diretamente entre as palavras, independentemente de uma tarefa de PLN específica (JURAVSKY e MARTIN, 2020). A metodologia consiste em estimar a similaridade entre pares de palavras calculada pelo modelo e compará-la à percepção humana, tomando como referência um conjunto de testes manualmente anotado (SCHNABEL *et al.*, 2015; GLADKOVA e DROZD, 2016; WANG *et al.*, 2019). Dessa forma, a métrica para a avaliação intrínseca consiste em, dado um par de termos, calcular a correlação entre a média dos índices de similaridade atribuídos por anotadores humanos e a similaridade cosseno obtida pelos modelos vetoriais (BARONI *et al.*, 2014; SCHNABEL *et al.*, 2015; CAMACHO-COLLADOS e PILEHVAR, 2018).

Portanto, para viabilizar o cálculo dessa correlação, é necessário dispor de um conjunto de dados de avaliação (*gold standard*) contendo os pares de termos relevantes para o domínio de O&G, e seus respectivos índices de similaridade baseados na percepção humana dos anotadores. Apesar de inúmeros trabalhos recentes para oferecer dados anotados para avaliação intrínseca (MIKOLOV *et al.*, 2013; ROGERS *et al.*, 2019), inclusive para o idioma português (RODRIGUES *et al.*, 2016, RODRIGUES e BRANCO, 2018), esses estudos são predominantemente focados em aspectos gerais de linguagem e não se aplicam aos propósitos deste trabalho, que objetiva viabilizar avaliações que contemplem as particularidades semânticas dos modelos especializados no domínio de O&G.

Nesse sentido, após uma detalhada pesquisa bibliográfica, não encontramos *datasets* anotados que permitissem a execução de uma metodologia de avaliação intrínseca com foco no domínio de O&G em português. Portanto, no contexto desta tese, como produto de uma parceria entre o Centro de Pesquisas e Desenvolvimento da

Petrobras (Cenpes) e as Universidades PUC-RS e UFRGS, foi conduzida uma iniciativa para criação de um *dataset* para avaliação intrínseca no domínio, anotado por especialistas da indústria e da comunidade acadêmica. A metodologia para anotação, criação do conjunto de testes e cálculo das métricas é detalhada na Seção 4.2.1, e os resultados das avaliações são discutidos na Seção 4.2.2.

4.2.1 Metodologia

Considerando a inexistência de dados anotados públicos em português para a avaliação intrínseca dos modelos no domínio de O&G, foi criado um conjunto de testes (*gold standard*) contendo índices de similaridades semântica para 1500 pares de termos técnicos. Os dados foram manualmente anotados por especialistas da indústria e da academia, das áreas de geociências e engenharia do Petróleo, seguindo uma abordagem baseada no conceito de relação semântica (*semantic relatedness*) (ZHANG *et al.*, 2013).

Como principal fonte de dados para identificar os termos mais relevantes no domínio, utilizamos um tesauro de referência nessa área, publicado pela Universidade de Tulsa: o *Petroleum Abstracts' Exploration & Production Thesaurus*[45] (a partir de agora referenciado simplesmente como *tesauro de E&P*). Uma versão traduzida desse tesauro nos foi cedida pela Memória Técnica da Petrobras, elaborada por empregados da empresa, o que nos permitiu a utilização segura dos termos corretamente traduzidos para o português.

A partir do *tesauro de E&P*, selecionamos 1500 pares de termos relevantes para o domínio, para compor o conjunto de testes a ser anotado pelos especialistas. A fim de se estabelecer uma distribuição balanceada entre os pares, e para evitar que o *dataset* seja ocasionalmente composto apenas por palavras não relacionadas entre si, as relações existentes no tesauro foram usadas como critério para composição dos pares. Isto é, a escolha dos pares de palavras foi curada de tal forma a manter um balanceamento entre palavras *altamente relacionadas*, palavras *parcialmente relacionadas*, e palavras *não relacionadas*. Em suma, o *tesauro de E&P* é essencialmente composto por três tipos de relações principais: *RELATED* (relações fracas não-explicitadas), *BROADER* e *NARROWER* (ambos indicando relações fortes de hiperonímia e hiponímia, respectivamente). Portanto, pares contendo palavras altamente relacionadas compõem 25% do conjunto de testes

[45] https://www.pa.utulsa.edu/products/tulsadatabase/thesaurus

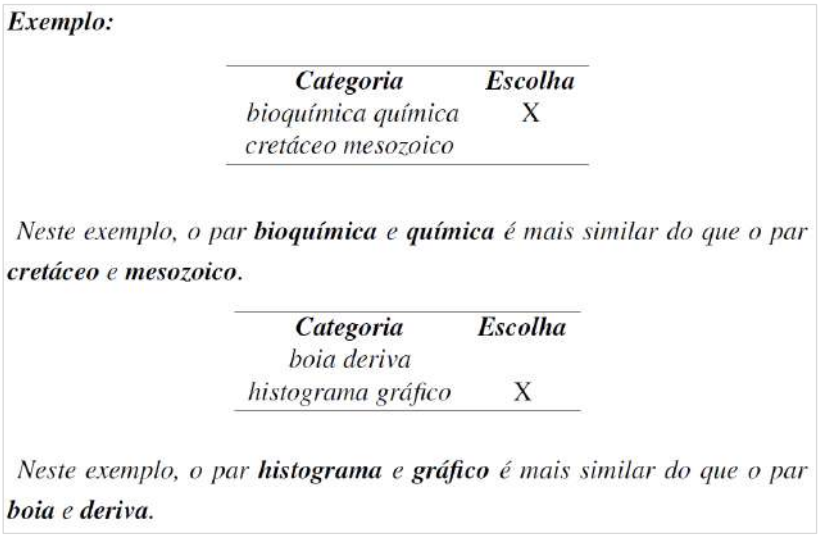
(aleatoriamente obtidos a partir de relações *BROADER* e *NARROWER*); 50% dos pares são de palavras parcialmente relacionadas (obtidas a partir de relações *RELATED* do tesauro); enquanto os demais 25% dos pares foram aleatoriamente compostos sem qualquer tipo de relação pré-estabelecida.

O processo de anotação dos dados envolveu a participação de 25 alunos de graduação do curso de geologia da UFRGS, um PhD em geociências, e dez especialistas do Centro de Pesquisa e Desenvolvimento da Petrobras (CENPES). Duas diferentes estratégias de anotação foram realizadas: *(i)* o anotador é apresentado a dois pares de termos, e deve selecionar qual o par contém os termos mais semanticamente relacionados entre si (método a partir de agora referenciado como *anotação binária de similaridade semântica*); e *(ii)* o anotador deve atribuir uma nota para representar o índice de similaridade entre os dois termos fornecidos, seguindo uma escala Likert de 1 a 7 (LIKERT, 1932) (método referenciado como *anotação Likert*).

O método *(i)*, referente ao processo de *anotação binária de similaridade semântica*, seguiu a proposta apresentada no trabalho de BRUNI *et al.* (2014). Primeiramente, um subconjunto de 300 pares foi aleatoriamente selecionado a partir do conjunto amostral de 1500 identificados no *tesauro de E&P*. Cada um desses pares foram então individualmente combinados com outros 24 pares aleatoriamente selecionados desse mesmo subconjunto, de forma que cada exemplo (par de palavras) seria anotado um total de 24 vezes. Os anotadores foram apresentados a esses conjuntos de dois pares de termos e orientados a selecionar qual deles contém os termos mais semanticamente relacionados entre si, conforme exemplo ilustrado na Figura 4.1. Cada vez que um par é selecionado, ele incrementa um ponto. Ao final do processo de anotação, cada par possui um *score* que varia de 0 a 24, conforme o número de vezes que tiver sido selecionado como o mais relevante. Desta forma, assume-se que os pares contendo as maiores pontuações possuem os pares de termos mais semanticamente relacionados entre si. Por fim, foram realizadas um total de 3600 comparações par-a-par (300 pares combinados entre si 24vezes, excluídas as duplicidades: $300 * 24/2$).

A *anotação Likert* (método *(ii)*), seguiu uma abordagem apresentada no trabalho de AGUIRRE *et al*. (2009). Para isto, todos os 1500 pares foram individualmente avaliados por uma equipe de anotadores, que foram orientados a atribuir uma nota conforme sua percepção de similaridade entre o par de termos, em uma escala entre 1 e 7 (onde 7 representa o valor máximo de similaridade). Nessa tarefa, cada *dataset* contendo

os 1500 pares foi anotado por três diferentes anotadores: um estudante do curso de Geologia da UFRGS, um P.hD em Geociências, e um especialista em engenharia de petróleo do CENPES. Em seguida, calculamos a estatística de concordância entre os anotadores usando a correlação Spearman, conforme apresentado na Tabela 4.1, obtendo o índice médio de *0,66*, que pode ser considerado ente moderada e forte, aceitável para os propósitos da tarefa. Por fim, calculamos a média aritmética entre os índices atribuídos pelos três anotadores para cada par de termos, para servir como referência para a cálculo da métrica intrínseca a ser comparada com as similaridades obtidas pelos modelos. O objetivo deste teste em comparação ao método *(i)* foi ampliar o tamanho do corpus de validação, considerando as limitações de esforço e disponibilidade de anotadores.



Exemplo:

Categoria	***Escolha***
bioquímica química	X
cretáceo mesozoico	

Neste exemplo, o par ***bioquímica*** *e* ***química*** *é mais similar do que o par* ***cretáceo*** *e* ***mesozoico****.*

Categoria	***Escolha***
boia deriva	
histograma gráfico	X

Neste exemplo, o par ***histograma*** *e* ***gráfico*** *é mais similar do que o par* ***boia*** *e* ***deriva****.*

Figura 4.1: Reprodução de um trecho das orientações enviadas aos anotadores sobre o processo de *anotação binária de similaridade semântica*, na qual cada anotador recebe dois pares de palavras e deve selecionar qual o par contém os termos mais semanticamente relacionados ente si

Tabela 4.1: Concordância entre os anotadores no método de anotação Likert
Fonte: GOMES *et al.* (2021)

Anotadores	Spearman
Anotadores 1 e 2	0.68
Anotadores 2 e 3	0.63
Anotadores 1 e 3	0.68
Média	**0.66**

A motivação para conduzir os dois processos de anotação em paralelo (métodos *(i)* e *(ii)*) consiste em garantir que os conjuntos de dados resultantes possam manter uma adequada qualidade de avaliação das relações semânticas, permitindo analisar a correlação entre esses dois diferentes métodos de anotação. O método *(i)* baseia-se em

julgamentos mais rápidos e mais fáceis pelos anotadores (escolher uma opção entre dois exemplos), mas demanda um conjunto maior de dados anotados; enquanto o método *(ii)* é ligeiramente menos intuitivo e demanda um tempo maior para cada anotação, mas necessita de menos anotadores humanos e resulta em um maior número de pares anotados. Portanto, para garantir que os dois conjuntos de dados anotados estavam correlacionados, comparamos o subconjunto dos 300 termos comuns aos dois *datasets*, obtendo um índice *Spearman* ρ de 0.86 entre os dois métodos, sugerindo uma forte correlação e indicando que o *dataset* baseado na *anotação Likert* era justo e adequado para ser utilizado como referência para o cálculo das métricas de avaliação intrínsecas.

Portanto, optamos por efetivamente utilizar o *dataset Likert* para calcular a métrica de avaliação intrínseca. Para cada par de termos, os índices de similaridade cosseno entre os seus vetores calculados pelos modelos foram correlacionados à média de seus respectivos índices anotados no *dataset Likert*. Seguindo BARONI *et al.* (2014), a correlação foi calculada utilizando o coeficiente de correlação ordinal de Spearman (*Spearman's rank correlation coefficient*, ou *Spearman* ρ)[46]. Além dessa correlação, nós também calculamos um índice de *cobertura* (*coverage*), *i.e.*, a quantidade dos termos que estão efetivamente presentes nos modelos vetoriais dividido pelo número total de termos do teste. A cobertura é um importante requisito para mensurar quanto do vocabulário de domínio está presente nos modelos, o que representa um aspecto crucial para o adequado funcionamento dos algoritmos de PLN especializados.

Por fim, de maneira a permitir um adequado ranqueamento dos resultados comparativamente entre os quatro modelos avaliados (**`baseline-O&G`**, **`skipgram-NILC`**, **`PetroVec-O&G`** e **`PetroVec-hybrid`**), é importante estabelecer uma métrica que considere tanto a *correlação* como a *cobertura* simultaneamente, uma vez que o melhor modelo deve ser capaz de capturar propriedades de similaridade semânticas e também garantir a inclusão da terminologia necessária ao domínio. Dessa forma, para agregar esses dois critérios em uma única métrica, também calculamos a *média harmônica* (*harmonic mean*)[47] entre a correlação e a cobertura. A justificativa foi simular o comportamento da amplamente utilizada métrica *F1*. Os resultados completos das métricas de avaliação intrínseca são apresentados a seguir, na Seção 4.2.2.

[46] Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_379
[47] Harmonic Mean. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_176

4.2.2 Resultados

A fim de avaliar a capacidade dos modelos em capturar propriedades semânticas do domínio a partir do corpus, a métrica de avaliação intrínseca foi calculada considerando cada um dos 1500 pares de termos, comparando a correlação entre os índices atribuídos pelos anotadores humanos e os scores de similaridade cosseno obtidos pelos diferentes modelos. A Tabela 4.2 apresenta os índices de correlação (*Spearman* ρ), cobertura e média harmônica, obtidos para cada um dos modelos.

Tabela 4.2: Resultados para a avaliação intrínseca considerando os diferentes modelos (*melhores resultados estão grifados em negrito*)
Fonte: GOMES *et al.* (2021)

Modelo	Arquitetura	*Spearman* (ρ)	Cobertura	Média harmônica
`baseline-O&G`	FastText	0,43	0,89	0,58
	Word2vec	0,51	**1,00**	0,68
`skipgram-NILC`	FastText	0,57	0,94	0,71
	Word2vec	0,48	0,94	0,64
`PetroVec-O&G`	FastText	0,56	**1,00**	0,72
	Word2vec	0,61	**1,00**	0,76
`PetroVec-Hybrid`	FastText	0,61	**1,00**	0,76
	Word2vec	***0,65***	**1,00**	**0,79**

Os resultados demonstram que os modelos baseados em corpora genéricos são inferiores aos modelos baseados em corpora específicos para o critério de cobertura, conforme esperado. Os resultados também confirmam que os modelos treinados em corpora maiores tendem a obter melhores resultados, o que pode ser observado pela superioridade do modelo **`PetroVec-O&G`** em relação ao modelo **`baseline-O&G`** (este último, treinado em um corpus significativamente menor). Entretanto, é interessante destacar que, nessa avaliação intrínseca, considerando os modelos especializados no domínio, o algoritmo FastText obteve desempenho ligeiramente inferior aos modelos baseados em Word2Vec. Curiosamente, o oposto foi observado para a avaliação extrínsecas com a tarefa de NER (apresentada na próxima seção), em que os modelos FastText obtiveram desempenho superior. Uma possível explicação para que esse cenário não ocorra no modelo genérico é o fato de que o modelo **`skipgram-NILC`** é distribuído pelo NILC em um formato intercambiável que perde algumas características da arquitetura FastText e que não faz pleno uso dos recursos de *n-grams*, fazendo-o atuar de forma mais parecida com um modelo word2vec na prática.

Adicionalmente, para avaliar se as diferenças encontradas nos coeficientes de correlação foram estatisticamente significantes, seguindo FARUQUI *et al.* (2016) e SHALABY e ZADROZNY (2017), calculamos o *Steiger's Z Test* (STEIGER, 1980) considerando as combinações de pares de modelos. Os resultados demonstram que o **`PetroVec-O&G`** usando Word2vec é significativamente melhor que **`baseline-O&G`** e **`skipgram-NILC`** (usando $alpha = 0{,}05$) para os dois algoritmos de vetorização. O melhor resultado geral é obtido pelo **`PetroVec-hybrid`** usando o algoritmo **Word2vec**, que obteve desempenho significativamente melhor que os demais modelos. A Tabela 4.3 ilustra os resultados completos para o teste *Steiger* entre os pares de modelos, enquanto a Figura 4.2 ilustra um gráfico comparativo dos resultados obtidos para os diferentes modelos, considerando a média harmônica.

Por fim, FARUQUI *et al.* (2016) destaca que o uso de tarefas de similaridade de palavras como critério para a avaliação intrínseca apresenta algumas limitações e recomenda que os modelos vetoriais sejam avaliados em tarefas aplicadas, *i.e.*, acompanhados também de avaliações extrínsecas. Portanto, para oferecer uma avaliação adequada dos modelos **`PetroVec`**, tais análises são abordadas na próxima seção.

Tabela 4.3: Resultados completos para o *Steiger's test* comparando os diferentes modelos. O nome do melhor modelo é apresentado em cada célula onde *p-value* $< 0{,}05$, e = para os demais casos. *O modelo com melhor resultado geral é grifado em negrito.*

	baseline-O&G Word2Vec	***baseline-O&G FastText***	***skipgram-NILC Word2Vec***	***skipgram-NILC FastText***	***PetroVec-O&G Word2Vec***	***PetroVec-O&G FastText***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid FastText***
baseline-O&G Word2Vec	=	baseline-O&G Word2Vec	=	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	***PetroVec-Hybrid Word2Vec***	PetroVec-Hybrid FastText
baseline-O&G FastText	baseline-O&G Word2Vec	=	=	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	***PetroVec-Hybrid Word2Vec***	PetroVec-Hybrid FastText
skipgram-NILC Word2Vec	=	=	=	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	***PetroVec-Hybrid Word2Vec***	PetroVec-Hybrid FastText
skipgram-NILC FastText	skipgram-NILC FastText	skipgram-NILC FastText	skipgram-NILC FastText	=	PetroVec-O&G Word2Vec	=	***PetroVec-Hybrid Word2Vec***	PetroVec-Hybrid FastText
PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	=	PetroVec-O&G Word2Vec	***PetroVec-Hybrid Word2Vec***	=
PetroVec-O&G FastText	PetroVec-O&G FastText	PetroVec-O&G FastText	PetroVec-O&G FastText	=	PetroVec-O&G Word2Vec	=	***PetroVec-Hybrid Word2Vec***	PetroVec-O&G FastText
PetroVec-Hybrid Word2Vec	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid Word2Vec***	=	***PetroVec-Hybrid Word2Vec***
PetroVec-Hybrid FastText	PetroVec-Hybrid FastText	PetroVec-Hybrid FastText	PetroVec-Hybrid FastText	PetroVec-Hybrid FastText	=	PetroVec-O&G FastText	***PetroVec-Hybrid Word2Vec***	=

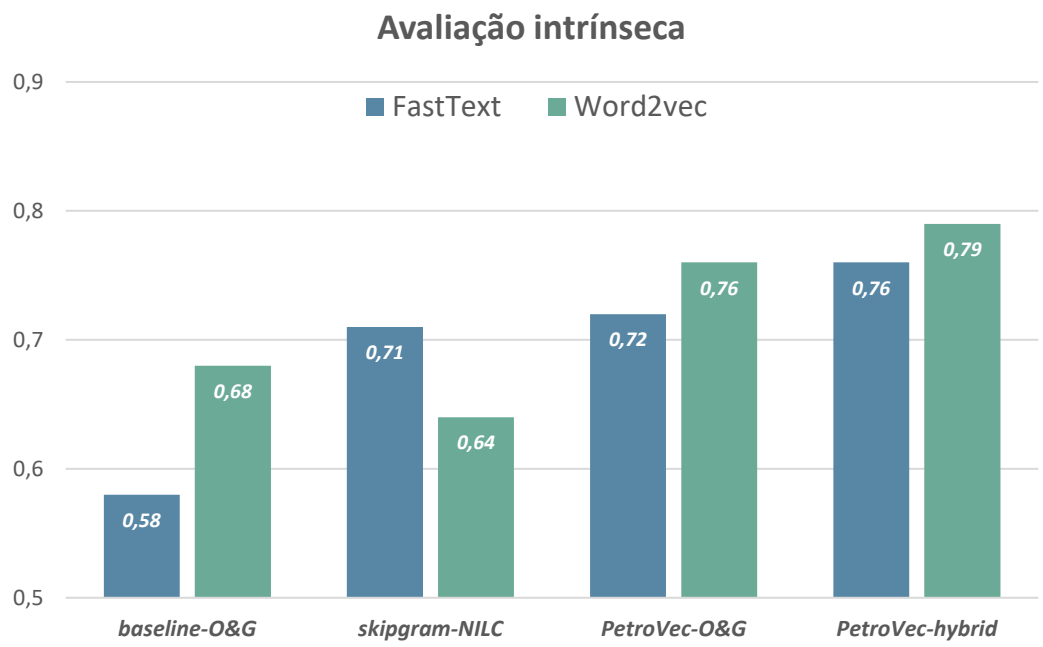


Figura 4.2: Comparativo do desempenho dos modelos, considerando a média harmônica *(com destaque para o range entre 0,5 e 0,9)*

4.3 Avaliação extrínseca

As avaliações extrínsecas objetivam medir a contribuição de um modelo de *word embedding* quando diretamente aplicado em algoritmos de PLN para uma tarefa específica, comumente focando em abordagens sintáticas (*e.g. POS tagging*, *parsing*) ou semânticas (*e.g.* classificação de textos, reconhecimento de entidades nomeadas, análise de sentimentos) (TURIAN *et al.*, 2010; SCHNABEL *et al.*, 2015; WANG *et al.*, 2019). Isto é, nas avaliações extrínsecas, os vetores são usados como representações das palavras na primeira camada de uma rede neural que implementa um algoritmo de PLN específico, e a qualidade dos modelos é avaliada a partir da melhora (ou piora) no desempenho do algoritmo nessa tarefa específica (JURAFSKY e MARTIN, 2020). De forma complementar às avaliações intrínsecas, as metodologias extrínsecas são importantes para melhor compreender a eficácia das técnicas de representação semântica de palavras quando diretamente utilizadas em aplicações do mundo real (PILEHVAR e CAMACHO-COLLADOS, 2020).

Neste trabalho, a avaliação extrínseca dos modelos **`PetroVec`** foi realizada em colaboração com as Universidades PUC-RS e UFRGS, baseada em uma tarefa aplicada de Reconhecimento de Entidades Nomeadas no domínio de Geociências. A metodologia para viabilização da métrica extrínseca é apresentada na Seção 4.3.1, e os resultados são discutidos na Seção 4.3.2.

4.3.1 Metodologia

Para viabilizar a tarefa de avaliação extrínseca considerando uma tarefa de REN, torna-se necessário dispor de um corpus contendo as anotações das entidades relevantes no domínio. Como a disciplina de Geociências está diretamente relacionada com a área de O&G, optamos por adotar uma versão revisada do GeoCorpus, originalmente apresentado por AMARAL (2017). Trata-se de um corpus anotado em português no domínio de Geologia para a tarefa de REN, contendo os identificadores de entidades geológicas relacionadas ao subdomínio *Bacia Sedimentar Brasileira*. O GeoCorpus é essencialmente composto por teses, dissertações, artigos e Boletins de Geociências da Petrobras, e contém nomes de rochas sedimentares, bacias sedimentares brasileiras, termos relacionados à Tectônica, Sedimentação e Magmatismo, e unidades estratigráficas (AMARAL, 2017). O GeoCorpus foi anteriormente utilizado como referência para um trabalho de REN desenvolvido por pesquisadores da PUC-RS em coautoria com o autor desta tese, cujos resultados foram recentemente publicados em um artigo apresentado na *Language Resources and Evaluation Conference (LREC 2020)* (CONSOLI *et al.*, 2020).

Nesse sentido, em um trabalho conduzido por pesquisadores da PUC-RS, o GeoCorpus foi detalhadamente revisado, a fim de melhor refletir os requisitos para a tarefa de avaliação. Portanto, foram providenciadas diversas correções para melhor adequação do corpus ao algoritmo de *machine learning*, como eliminação de categorias vazias e aninhadas, remoção de linhas duplicadas, correção de linhas incorretamente segmentadas e uma padronização geral das categorias. Além disso, o corpus foi incrementado com novas categorias específicas relevantes para o domínio, como por exemplo, a substituição da ampla e heterogênea categoria *OUTROS* por diversas categorias mais específicas, como *ROCHA MAGMÁTICA*. A base GeoCorpus revista e atualizada, assim como o relatório das principais modificações produzidas pela equipe de pesquisadores da PUC-RS, encontram-se disponíveis no repositório público **`PetroVec`**[48].

Portanto, a versão completa e revisada do GeoCorpus utilizada neste trabalho contempla 5275 sentenças, com 8933 entidades nomeadas, subdivididas em 30 categorias. A fim de garantir uma melhor diversidade léxica na condução da avaliação extrínseca, selecionamos as 10 categorias com a maior variedade e o maior número de entidades nomeadas. Essa escolha foi motivada pela tentativa de evitar que categorias

[48] https://github.com/Petroles/Petrovec/tree/master/GeoCorpus%20V3

menos representativas pudessem causar ruídos nos resultados, uma vez que classes com menor diversidade lexical poderiam ser mais facilmente identificadas pela rede, possivelmente causando *overfitting*. Nesse sentido, foram selecionadas as sentenças contendo pelo menos uma instância de entidade entre as 10 categorias mais frequentes, totalizando uma amostragem de 2978 sentenças, com 6578 entidades nomeadas. Uma tabela contendo o detalhamento completo das 30 categorias contempladas na revisão do GeoCorpus está disponível no material suplementar do artigo publicado pelo autor desta tese (GOMES *et al.*, 2021), e reproduzida em português no Apêndice A.1. Detalhes da composição das categorias usadas nesta avaliação extrínseca são apresentados na Tabela 4.5, que descreve os resultados individualmente para cada classe considerando o modelo que obteve o melhor desempenho.

Similarmente ao realizado na avaliação intrínseca e apresentado na Seção 4.2, na etapa de avaliação extrínseca também foram consideradas análises comparativas entre as variações dos modelos **`PetroVec`**, geradas a partir das diferentes composições de corpora, além do modelo genérico **`skipgram-NILC`** e do modelo preliminar **`baseline-O&G`**, que serviram como referência para as métricas.

A implementação do algoritmo de avaliação baseou-se em uma rede neural de estado-da-arte para a tarefa de REN fornecida por SANTOS *et al.* (2019), que foi alimentada pelos modelos vetoriais pré-treinados e as diferenças de desempenho entre os modelos foram apuradas. Essa rede neural, baseada na arquitetura FLAIR (AKBIK *et al.*, 2018), possui um módulo para receber o acoplamento dos modelos vetoriais de forma simplificada, tornando-a adequada para a realização dos testes extrínsecos, cujo objetivo principal consiste em avaliar comparativamente o impacto dos diferentes modelos de vetorização na tarefa, e não necessariamente os resultados do algoritmo de REN. As avaliações foram realizadas a partir de validação cruzada (*cross-validation*), com 10 subconjuntos amostrais gerados aleatoriamente para os dados de treinamento (cerca de 70% do corpus), teste (20%) e validação (10%), buscando preservar uma distribuição adequada para as instâncias de entidades.

4.3.2 Resultados

A Tabela 4.4 apresenta a média dos resultados para cada um dos oito modelos testados, considerando $k = 10$ (dez grupos de validação cruzada), contemplando os índices de precisão, revocação (*recall*) e F1. Os melhores resultados foram obtidos pelos

modelos **PetroVec-O&G** e **PetroVec-hybrid** (ambos usando FastText). Essa superioridade é atribuída principalmente pela maior revocação, consequência de uma cobertura mais ampla do vocabulário técnico contemplada no corpus de treinamento desses modelos. A Figura 4.3 ilustra um gráfico comparativo dos resultados obtidos pelos diferentes modelos, considerando a métrica F1.

Tabela 4.4: Resultados para a avaliação extrínseca de REN considerando os diferentes modelos (média 10-*fold cross-validation*), *melhores resultados por métrica estão grifados em negrito*
Fonte: GOMES *et al.* (2021)

Modelo	Arquitetura	Precisão	Revocação	F1
baseline-O&G	FastText	0,81	0,75	0,78
	Word2vec	0,78	0,69	0,74
skipgram-NILC	FastText	**0,83**	0,82	0,83
	Word2vec	0,80	0,80	0,80
PetroVec-O&G	FastText	**0,83**	**0,89**	**0,86**
	Word2vec	0,82	0,81	0,81
PetroVec-Hybrid	FastText	**0,83**	**0,89**	**0,86**
	Word2vec	0,82	0,81	0,82

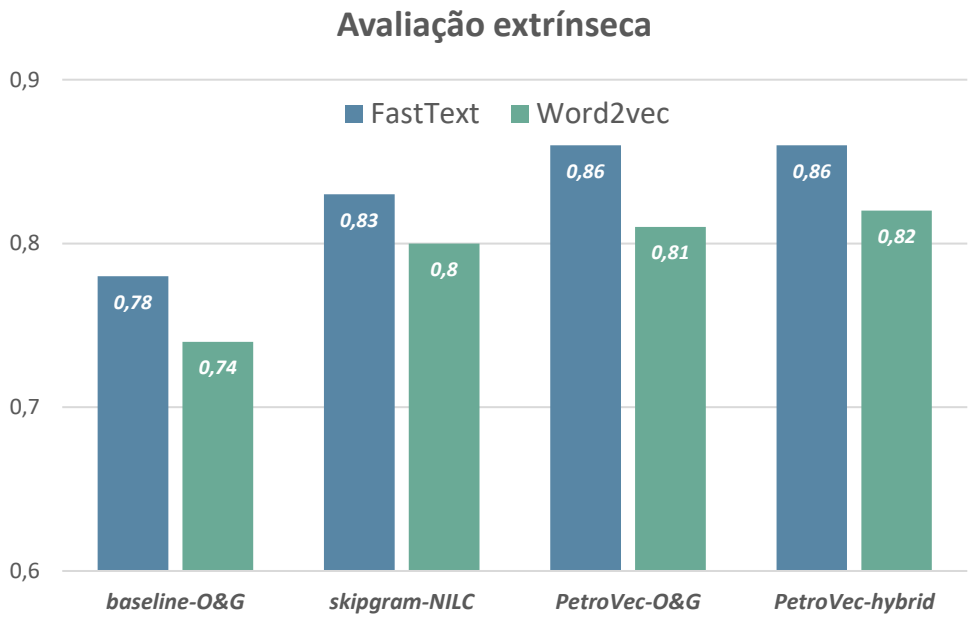


Figura 4.3: Comparativo do desempenho dos modelos para avaliação extrínseca de REN, considerando a métrica F1 *(com destaque para o range entre 0,6 e 0,9)*

Em relação ao corpus, uma comparação entre os modelos **PetroVec** e **skipgram-NILC** demonstra que a utilização de corpora específico no domínio permitiu ao modelo especializado obter melhores métricas em relação à cobertura de vocabulário técnico, mesmo diante de uma menor disponibilidade de dados para o treinamento. Isto é, os modelos **PetroVec** foram capazes de obter resultados superiores em relação ao

modelo genérico, mesmo tendo sido treinados em um corpus cerca de uma ordem de magnitude menor – *i.e.*, o corpus de treinamento utilizado pelo modelo genérico **skipgram-NILC** é cerca de 16 vezes maior do que o nosso *corpus especializado* de O&G, conforme anteriormente ilustrado na Figura 3.3. Não obstante, é relevante destacar que o tamanho do corpus também desempenha papel importante – o que pode ser comprovado comparando o modelo **PetroVec-O&G** com a versão preliminar **baseline-O&G**, treinada em um corpus significativamente menor. Os bons resultados do modelo **skipgram-NILC**, particularmente na métrica de precisão, podem ser também atribuídos ao tamanho do seu corpus, uma vez que, mesmo sem cobertura significativa para vocabulário técnico específico do domínio, o modelo foi capaz de generalizar propriedades semânticas do idioma português, permitindo inferir corretamente cerca de 80% dos casos. Em contrapartida, não observamos ganhos significativos de desempenho ao adicionar conteúdo genérico ao corpus de treinamento específico, conforme pode ser concluído a partir das métricas muito similares entre os modelos **PetroVec-O&G** e **PetroVec-hybrid**.

Analisando especificamente as duas abordagens para os algoritmos de treinamento (Word2Vec e FastText), é possível observar que o FastText apresentou resultados consistentemente superiores em relação às suas respectivas contrapartes Word2Vec, tendo sido capaz de incrementar as métricas de *revocação* para os modelos **PetroVec-O&G** e **PetroVec-hybrid** em cerca de oito pontos percentuais. Para esta tarefa, portanto, os resultados sugerem que a capacidade da arquitetura FastText em lidar com informações de subpalavras permite melhor capturar propriedades linguísticas específicas do corpus especializado.

Em uma análise mais detalhada, a Tabela 4.5 apresenta a média de resultados individualizada por categoria, considerando um dos modelos que obtiveram melhor desempenho na avaliação extrínseca (*i.e.*, **PetroVec-O&G** FastText). O modelo obteve bons resultados para as métricas de precisão e revocação de maneira majoritariamente uniforme entre as categorias, com uma particular exceção: a métrica de revocação para a categoria *IDADE*, equivalente a 0,97 e quase 0,1 pontos acima da média, enquanto sua precisão foi de 0,76 e quase 0,1 abaixo da média. Analisando particularmente as entidades dessa categoria, observamos um padrão morfológico, cujas entidades comumente terminam em um dos tetragramas: *"iano"/"iana"* (*e.g.*, Artinskiano); ou *"eano"/"eana"* (*e.g.*, Zancleano). Acreditamos que a característica do FastText em utilizar informações

de subpalavras pode ter contribuído para esses resultados, permitindo ao modelo reconhecer essa categoria com uma melhor capacidade de revocação, porém possivelmente ocasionando um efeito adverso de classificar incorretamente outras palavras que contenham os mesmos tetragramas. De toda forma, considerando as características gerais de morfologia do vocabulário de O&G, acreditamos que essa representação de palavras a partir de sequências de caracteres (*n-grams*) ainda seja útil para permitir uma correta identificação de termos técnicos raros particulares ao domínio e para lidar com palavras fora do vocabulário, permitindo identificar morfemas comuns e tirando benefício da estrutura morfológica do idioma português e suas diversas inflexões nominais e verbais (BOJANOVSKI *et al.*, 2017).

Tabela 4.5: Categorias, número de instâncias e resultados REN para o modelo `PetroVec-O&G`
Fonte: GOMES *et al.* (2021)

Categoria	Instâncias	Precisão	Revocação	F1
Rochas				
Rocha Sedimentar Siliciclástica	1098	0,85	0,88	0,86
Rocha Magmática	580	0,86	0,95	0,91
Rocha Metamórfica	377	0,83	0,86	0,85
Rocha Sedimentar Carbonática	355	0,79	0,87	0,83
Tempo				
Idade	796	0,76	0,97	0,85
Período	712	0,92	0,92	0,92
Época	686	0,90	0,90	0,90
Elementos Estatigráficos				
Unidade Estatigráfica	763	0,81	0,86	0,83
Bacia Sedimentar	551	0,86	0,88	0,86
Lugares				
Contexto Geológico de Bacia	660	0,75	0,83	0,79
Total	**6578**	**0,83**	**0,89**	**0,86**

De maneira similar ao realizado para as avaliações intrínsecas, a fim de demonstrar a significância estatística dos resultados através de testes-t, primeiramente calculamos o teste de normalidade Kolmogorov-Smirnov[49] em cada série dos resultados. As análises confirmam que nossos dados não diferem significativamente daqueles que são normalmente distribuídos. Em seguida, os testes-t subsequentes demonstraram que os ganhos em revocação obtidos pelo **`PetroVec-O&G`** comparados ao **`skipgram-NILC`**

[49] Kolmogorov–Smirnov Test. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_214

são estatisticamente significantes (p-value = 0,001). Os resultados completos dos testes-t entre os pares de modelos são apresentados na Tabela 4.6.

Tabela 4.6: Resultados *t-test* de significância estatística comparando os pares de modelos. O nome do melhor modelo é apresentado em cada célula onde *p-value* $< 0,05$, e = para os demais casos. *O modelo com melhor resultado geral é grifado em negrito.*

	baseline-O&G Word2Vec	***baseline-O&G FastText***	***skipgram-NILC Word2Vec***	***skipgram-NILC FastText***	***PetroVec-O&G Word2Vec***	***PetroVec-O&G FastText***	***PetroVec-Hybrid Word2Vec***	***PetroVec-Hybrid FastText***
baseline-O&G Word2Vec	=	*baseline-O&G FastText*	*skipgram-NILC Word2Vec*	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	*PetroVec-Hybrid Word2Vec*	***PetroVec-Hybrid FastText***
baseline-O&G FastText	*baseline-O&G FastText*	=	*skipgram-NILC Word2Vec*	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	*PetroVec-Hybrid Word2Vec*	***PetroVec-Hybrid FastText***
skipgram-NILC Word2Vec	*skipgram-NILC Word2Vec*	*skipgram-NILC Word2Vec*	=	skipgram-NILC FastText	PetroVec-O&G Word2Vec	PetroVec-O&G FastText	*PetroVec-Hybrid Word2Vec*	***PetroVec-Hybrid FastText***
skipgram-NILC FastText	skipgram-NILC FastText	skipgram-NILC FastText	skipgram-NILC FastText	=	skipgram-NILC FastText	*PetroVec-O&G FastText*	=	***PetroVec-Hybrid FastText***
PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	PetroVec-O&G Word2Vec	skipgram-NILC FastText	=	PetroVec-O&G FastText	=	***PetroVec-Hybrid FastText***
PetroVec-O&G FastText	PetroVec-O&G FastText	PetroVec-O&G FastText	PetroVec-O&G FastText	*PetroVec-O&G FastText*	=	=	*PetroVec-Hybrid Word2Vec*	***PetroVec-O&G FastText***
PetroVec-Hybrid Word2Vec	*PetroVec-Hybrid Word2Vec*	*PetroVec-Hybrid Word2Vec*	*PetroVec-Hybrid Word2Vec*	PetroVec-O&G FastText	=	*PetroVec-Hybrid Word2Vec*	=	***PetroVec-Hybrid FastText***
PetroVec-Hybrid FastText	***PetroVec-Hybrid FastText***	***PetroVec-Hybrid FastText***	***PetroVec-Hybrid FastText***	***PetroVec-Hybrid FastText***	***PetroVec-Hybrid FastText***	***PetroVec-O&G FastText***	***PetroVec-Hybrid FastText***	=

4.4 Avaliações qualitativas

De maneira complementar às avaliações quantitativas intrínsecas e extrínsecas apresentadas nas seções anteriores, conduzimos também um conjunto adicional de experimentos baseados em análises qualitativas, com o objetivo de oferecer uma cobertura consistente de métodos para avaliar a qualidade dos modelos **`PetroVec`**. Essas avaliações qualitativas têm por objetivo verificar a qualidade de representação semântica dos modelos **`PetroVec`**, com foco principal no vocabulário técnico do domínio de O&G. Portanto, as análises foram realizadas utilizando o modelo **`PetroVec-O&G`**, treinado exclusivamente a partir do *corpus especializado*, e contemplam abordagens baseadas em *analogias de palavras*, *coerência de espaço semântico* e *categorização de conceitos* a partir do espaço vetorial (TURIAN *et al.*, 2010; SCHNABEL *et al.*, 2015; MUNEEB *et al.*, 2015). Nas subseções a seguir, apresentamos uma detalhada descrição das análises realizadas e uma discussão sobre respectivos resultados obtidos.

4.4.1 Analogia de Palavras

As avaliações qualitativas por analogias de palavras objetivam encontrar um termo *y* para um dado termo *x*, de modo a relação *x:y* se assemelhe a uma relação de referência *a:b* (SCHNABEL *et al.*, 2015). MIKOLOV *et al.* (2013) demonstraram a existência de certas regularidades linguísticas capturadas no espaço vetorial, de forma que modelos de vetorização de palavras podem resolver algumas relações semânticas simples entre termos. Essas analogias podem ser expressas da forma "*a* está para *b* assim como *x* está para *y*", como no exemplo: *"*França *está para* Paris, *assim como* Alemanha *está para* **?***",* onde '**?**' é oculto e para o qual se espera que o modelo resolva a analogia encontrando a palavra *"Berlim"* (MIKOLOV *et al.* ,2013). Essas analogias semânticas entre termos podem ser resolvidas calculando a similaridade cosseno a partir da diferença entre os vetores das palavras de referência (Equação(4.1) (PADARIAN *et al.*, 2019):

$$\frac{(v_b - v_a)^{\intercal} \cdot (v_y - v_x)}{\|v_b - v_a\| \, \|v_y - v_x\|} \qquad (4.1)$$

Portanto, neste trabalho exploramos operações de analogias de palavras para investigar a capacidade dos modelos **`PetroVec`** em codificar relações linguísticas entre termos do vocabulário técnico de O&G, capturadas automaticamente e de forma não-supervisionada a partir dos corpora de treinamento. Primeiramente, utilizando como referência o conteúdo em português disponibilizado pelo Dicionário do Petróleo (FERNÁNDEZ *et al.*, 2009) [50] , selecionamos três categorias de termos técnicos relacionados a subdomínios de O&G: *(i)* *GEOLOGIA*, *(ii)* *PROFISSÕES TÉCNICAS*, e *(iii)* *UNIDADES DE MEDIDAS E INSTRUMENTOS*. Para cada categoria, fornecemos um par de palavras para servir como relação de referência (correspondendo ao parâmetro *a:b*), e utilizamos o modelo **`PetroVec-O&G`** para calcular as operações de analogias semânticas para uma lista de termos técnicos (*x*) previamente selecionados. Isto é, calculamos as operações de analogia, conforme descrito na Equação (4.1 , obtendo uma palavra *y* para cada exemplo de *x* que atenda a relação de amostra *a:b*. Por exemplo, considerando a categoria *PROFISSÕES TÉCNICAS*, dada a relação de amostra *"geólogo* está para *geologia"* (*a:b*), calculamos as operações de analogia para cada um dos termos de referência (*x*),

[50] Dicionário do Petróleo. Disponível em: http://dicionariodopetroleo.com.br/. Acesso em 01/05/2020.

obtendo como resultado a palavra mais próxima (*y*) ao vetor resultante calculado pelo modelo **PetroVec-O&G**, encontrando resultados como *"engenheiro"* está para *"engenharia"*. Em seguida, as mesmas operações de analogias semânticas foram calculadas usando o modelo genérico **skipgram-NILC.** Os resultados obtidos para os diferentes modelos são comparativamente apresentados na Tabela 4.7.

É possível observar que o modelo **PetroVec-O&G** foi capaz de identificar corretamente os termos resultantes das operações de analogia, conforme esperado, mantendo a coerência das relações semânticas no contexto do domínio de O&G. Em contraste, apesar do modelo **skipgram-NILC** apresentar bons resultados para alguns exemplos da categoria *PROFISSÕES TÉCNICAS*, possivelmente em função de um aspecto generalista mais amplo desses conceitos, falhou em todas as demais categorias. Além disso, uma análise mais abrangente dos resultados sugere que o modelo especializado **PetroVec-O&G** foi capaz de capturar automaticamente propriedades sintáticas a partir do corpus de domínio, conforme observado na sequência de termos contendo o mesmo sufixo (*i.e.*, o sufixo *'metro'* em '*manômetro*', '*densímetro*' e '*porosímetro*' da categoria *MEDIDAS E INSTRUMENTOS*), além de outras propriedades semânticas mais complexas e que não compartilham quaisquer radicais ou sufixos (*e.g.*, '*bureta*' e '*bússola*' da mesma categoria).

Tabela 4.7: Operações de analogias semânticas obtidas pelos modelos skipgram-NILC e PetroVec, a partir do exemplo de referência, na forma: *a* está para *b* como *x* está para *?*

Geologia *silte : siltito*	**Profissões Técnicas** *geologo : geologia*	**Medidas e Instrumentos** *temperatura : termometro*
PetroVec		
folhelho : arenito	geofisico : geofisica	pressao : manometro
argila : argilito	engenheiro : engenharia	volume : bureta
calcario : dolomito	quimico : quimica	direcao : bussola
sal : evaporito	gestor : gestao	porosidade : porosimetro
feldspato : hornblenda	oceanografo : oceanografia	densidade : densimetro
carbonato : anidrita	enfermeiro : enfermagem	corrente : multimetro
skipgram:NILC		
folhelho : lagerstätte	geofisico : geofísica	pressao : apetite
argila : lacados	engenheiro : engenharia	volume : negócio
calcario : lobeinstein	quimico : química	direcao : c-leg
sal : meloa	gestor : gerência	porosidade : dielétrico
feldspato : feldspatos	oceanografo : oceanografia	densidade : dólar-turismo
carbonato : hidróxido	enfermeiro : odontologia	corrente - modismo

Adicionalmente, é relevante ressaltar que, no decorrer dos experimentos, percebemos que os modelos **`PetroVec`** foram capazes de relacionar alguns termos do vocabulário técnico em português com sua correspondente tradução para o inglês, utilizando apenas operações de analogias semânticas. Esse comportamento pôde ser observado mesmo considerando que os corpora usados no treinamento foram compostos prioritariamente por documentos originalmente obtidos em português, não contendo nenhum tipo de conteúdo paralelo bilíngue explicitamente projetado. Portanto, as relações semânticas evidenciadas pelas analogias bilíngues foram capturadas automaticamente de maneira não-supervisionada a partir do corpus de domínio. A Tabela 4.8 descreve alguns experimentos com analogias bilíngues corretamente identificadas pelo modelo, tendo sido fornecido apenas o exemplo de referência *'reservatorio' está para 'reservoir'* (*a:b*). Adicionalmente, seguindo o experimento originalmente apresentado por MIKOLOV *et al.* (2013a), a Figura 4.4 apresenta uma projeção bidimensional dos vetores dos termos contemplados nas operações de analogias, usando análise de componentes principais (*Principal Component Analysis*, PCA) (JOLLIFFE, 1986) para redução de dimensionalidade.

Tabela 4.8: Exemplos de relações bilíngues corretamente identificadas pelas operações de analogias semânticas, dado o exemplo de referência *'reservatorio:reservoir'*
Fonte: GOMES *et al.* (2021)

Analogias semânticas português-inglês Referência: *reservatorio : reservoir*	
exploracao : *exploration*	producao : *production*
perfuracao : *drilling*	sismica : *seismic*
campo : *field*	bacia : *basin*
plataforma : *platform*	oleo : *oil*
hidrocarboneto : *hydrocarbon*	combustivel : *fuel*
duto : *pipe*	rocha : *rock*
falha : *fault*	poco : *wells*
porosidade : *porosity*	permeabilidade : *permeability*
viscosidade : *viscosity*	pressao : *pressure*
arenito : *sandstone*	sal : *salt*
sedimento : *sediment*	brasil : *brazil*

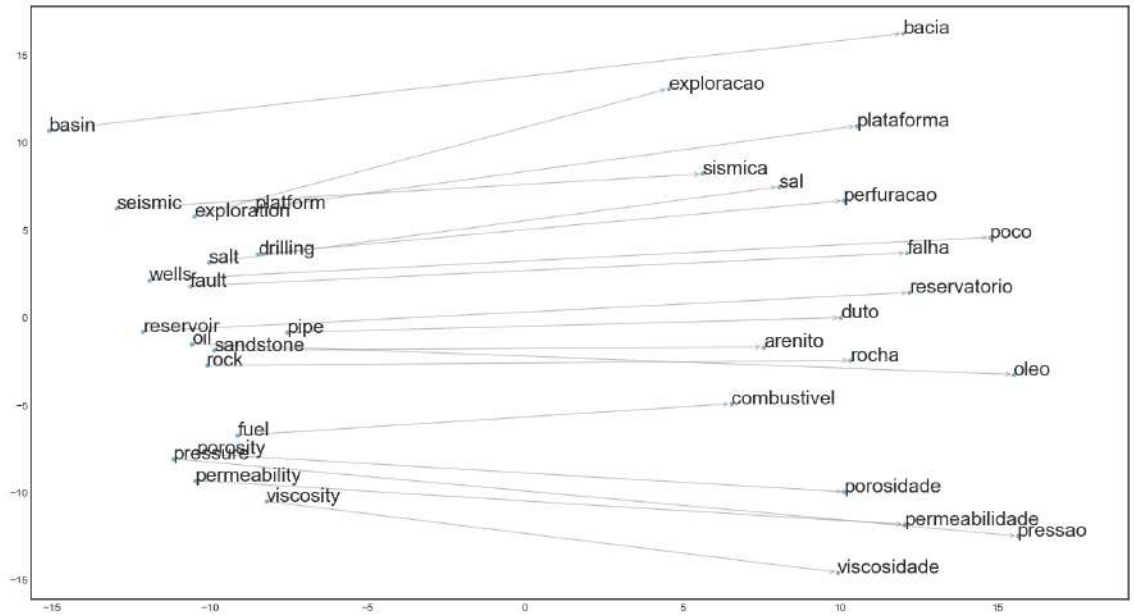


Figura 4.4: Projeção PCA bidimensional das analogias envolvendo traduções português-inglês para uma amostra de termos técnicos selecionados.
Fonte: GOMES *et al.* (2021)

Cabe ressaltar que o comportamento observado para as analogias bilíngues não é extensivo a todos os possíveis casos do vocabulário, uma vez que os modelos **`PetroVec`** não se propõem a ser uma ferramenta de tradução automática. Não obstante, conseguimos replicar com sucesso esse comportamento para um número significativo de exemplos a partir do vocabulário técnico especializado de O&G, o que pode sugerir um indicador positivo da capacidade de generalização dos modelos. Nesse sentido, realizamos análises complementares para investigar cenários adicionais em que as analogias de tradução se mantêm coerentes. Seguindo ROGERS *et al.* (2017) e NEWMAN-GRIFFIS *et al.* (2017), ampliamos as condições de similaridade da palavra-alvo de forma a admitir a existência de múltiplas palavras corretas, listando os *top*-3 termos mais próximos do vetor resultante das operações de analogia, permitindo assim a análise de uma região de vizinhança mais abrangente. A Tabela 4.9 apresenta os resultados para esses experimentos, considerando a mesma relação de referência anterior (*'reservatorio'*;*'reservoir'*), e um conjunto complementar de termos de pesquisa. É possível observar que para a maioria dos casos o modelo **`PetroVec-O&G`** retornou corretamente o termo esperado entre os três vizinhos mais próximos. Além disso, para os demais exemplos, apesar do termo esperado não constar entre os *top*-3, nota-se que as palavras na região de vizinhança retornadas pelo modelo se mostram semanticamente relacionadas ao conceito de tradução esperado (*e.g.*, considerando o termo *prospeccao*, o modelo retorna elementos semanticamente relacionados ao conceito esperado para a

tradução resultante, como *seismic*, *exploration* e *exploratory*). Adicionalmente, em casos específicos, o modelo apresenta mais de uma resposta considerada satisfatória para a analogia, conforme preconizado por NEWMAN-GRIFFIS *et al.* (2017), inclusive contemplando conceitos polissêmicos como a tradução de '*perfil'* para '*profile*' e também para '*logs*' (que seria adequada para perfis de poço, com tradução correspondente como *well logs*). Independentemente, essa propriedade observada no espaço semântico para as analogias de tradução sugere uma forte capacidade de generalização dos modelos para esses conceitos bilíngues, o que é particularmente útil para domínios globalizados como O&G, sendo uma característica crucial para uma série de tarefas aplicadas de PLN como busca semântica, sistemas de perguntas e respostas, plataformas conversacionais e tradução automática.

Tabela 4.9: Operações de analogias admitindo região de vizinhança expandida, considerando a relação de referência '*reservatorio : reservoir*' (*os resultados esperados estão grifados em negrito*) Fonte: GOMES *et al.* (2021)

Analogias semânticas português-inglês	
Referência: *reservatorio : reservoir*	
particula : *frsg*, *droplet*, ***particle***	completacao : *well*, ***completion***, *drilling*
batimetria : *depth*, *seismic*, *spatial*	poros : *melim*, *vuggy*, ***pore***
geofisica : *seismic*, *geopro*, *sulfabras*	cimentacao : *completion*, ***cementing, cementation***
asfalto : *speedy*, *brasivil*, ***asphalt***	navio : *moored*, ***ship***, *operations*
camada : *wiu*, ***layer***, *caniada*	estimativa : ***estimation***, *proxy*, ***prediction***
risco : ***risk***, *probability*, *attractiveness*	perfil : *perl*, ***profile, logs***
planta : ***plant***, ***facility***, *microplanta*	prospeccao : *seismic*, *exploration*, *exploratory*
bomba : *centrifugal*, ***pump***, *microvalve*	condutividade : *permeability*, *density*, ***conductivity***

Para confirmar que os resultados obtidos pelas analogias bilíngues não se tratam de um artefato de corpus, nós analisamos a frequência com que os pares de termos contemplados nas analogias co-ocorrem no corpus de treinamento. Portanto, considerando os pares de palavras anteriormente apresentados na Tabela 4.8, calculamos a frequência individual de cada termo em relação à frequência de co-ocorrência dos pares de palavras, considerando uma janela de contexto deslizante de tamanho 5 (que reflete o mesmo valor de hiperparâmetro para a janela de contexto usada no treinamento dos vetores). Os resultados demonstraram que a co-ocorrência dos pares de termos dentro de uma mesma janela de contexto é muito rara em comparação à frequência individual de cada palavra. Calculando o coeficiente Jaccard para os pares de palavras da analogia bilíngue, observamos um índice médio de apenas 0,1%. O maior índice encontrado para o par mais frequente é de apenas 0,44%, conforme apresentado em detalhes na Figura 4.5.

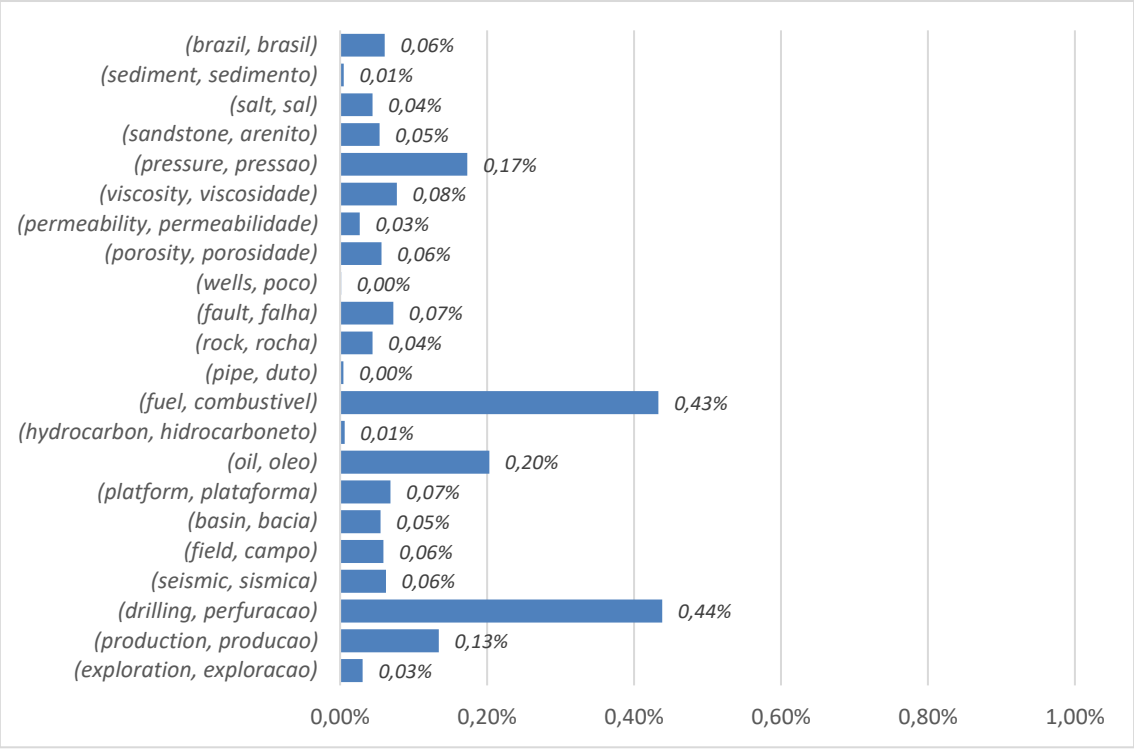


Figura 4.5 Frequência de co-ocorrência (índice Jaccard) dos pares de termos utilizados nas analogias semânticas bilingues, considerando janela de contexto de 5 palavras

Adicionalmente, seguindo a metodologia proposta por LINZEN (2016), realizamos um conjunto de análises complementares para validar as operações de analogias semânticas, avaliando se a qualidade dos resultados seria afetada ao inverter a direção das operações. Isto é, considerando a relação de referência anterior (*reservatorio:reservoir*), ao inverter a direção dessa relação (*reservoir:reservatorio*), é desejável que as analogias resultantes possam manter sua coerência semântica. Para esses experimentos, a partir de agora referenciados como *analogias reversas*, seguindo ROGERS *et al.* (2017) e NEWMAN-GRIFFIS *et al.* (2017), optamos por admitir múltiplas respostas corretas, listando os top-3 termos mais similares e igualmente permitindo analisar uma região de vizinhança mais ampla. Os resultados evidenciam que, apesar de uma leve queda de qualidade, as *analogias reversas* se mostraram majoritariamente consistentes em seus resultados, sendo possível observar os termos esperados corretamente listados na ampla maioria dos casos. Além disso, é relevante destacar que que a região de vizinhança obtida se mantém semanticamente consistente na maior parte dos experimentos. Os resultados completos são apresentados na Tabela 4.10.

De forma complementar, a mesma metodologia de analogias reversas foi também aplicada no conjunto de dados apresentados na Tabela 4.7, contemplando as

categorias *(i)* *GEOLOGIA*, *(ii)* *PROFISSÕES TÉCNICAS*, e *(iii)* UNIDADES DE MEDIDAS E INSTRUMENTOS. Porém, diferentemente do experimento anterior com as analogias bilíngues, neste caso foi possível observar uma leve queda de qualidade nos resultados. Entretanto, observamos que, para a ampla maioria dos exemplos, os termos esperados são corretamente listados entre os 5 primeiros resultados (13 entre os 18 exemplos analisados, uma taxa de sucesso de aproximadamente 72%). Além disso, nos demais casos, os resultados obtidos ainda estão semanticamente relacionados à categoria esperada, conforme apresentado na Tabela 4.11, o pode ser considerado como aceitável para indicar uma boa capacidade de generalização dos modelos, conforme os propósitos deste trabalho.

Tabela 4.10: Analogias reversas para os exemplos de pares de termos bilíngues, após inverter a direção de cada analogia (*os resultados esperados estão grifados em negrito*)

Analogias semânticas português-inglês Referência: *reservoir : reservatorio*	
exploration : *poco, pocos,* ***exploracao***	production: ***producao***, *erupcao, produzido*
drilling : *poco,* ***perfuracao***, *completacao*	seismic : ***sismica, sismico***, *sismicas*
field : ***campo***, *poco, furo*	basin : ***bacia***, *subbacia, bacias*
platform : ***plataforma***, *poco, monocoluna*	oil : ***oleo***, *condensado, petroleo*
hydrocarbon : ***hidrocarboneto***, *hidrocarbonetos, rocha*	fuel : ***combustivel***, *motor, ignicao*
pipe : ***duto***, *revestimento, falha*	rock : ***rocha***, *encaixante, canhoneios*
fault : ***falha***, *falhas, falhamento*	wells : *poco,* ***pocos***, *canhoneios*
porosity : ***porosidade***, *permeabilidade, rocha*	permeability : ***permeabilidade***, *porosidade, fratura*
viscosity : ***viscosidade***, *fluido, viscoso*	pressure : ***pressão***, *fluido, poco*
sandstone : ***arenito***, *folhelho, aquifero*	salt : ***sal***, *aquifero, diapiro*
sediment : ***sedimento***, *agua, solo*	brazil : *sul, bacia, sudoeste*

Tabela 4.11: Analogias reversas considerando os diferentes subdomínios relacionados a O&G (os resultados esperados estão grifados em negrito)

Geologia *silte : siltito*	**Profissões Técnicas** *geologo : geologia*	**Medidas e Instrumentos** *temperatura : termometro*
	PetroVec	
arenito : ***areia***, *granulometria,* ***areias***, ***argila***, *grossos*	geofisica : ***geofisico***, *geologos, geologa, consultora, depex*	manometro : ***pressao***, *vazao, saida, saturacao, velocidade*
argilito : *siltosa, siltica,* ***areia***, *silticas, calhaus*	engenharia : ***engenheiro***, *segen, estagiario, supervisor, consultora*	bureta : *saturacao, mistura, fracao, vazao, massa*
dolomito : *siltosa, siltica, lamosa, detriticas, caulinitas*	quimica : ***quimico***, *mota, quimicas, qufmica, medicamentosa*	bussola : *velocidade, inclinacao, magnetizacao, magnetico, rotacao*
evaporito : *terrigena, cascalho, infiltrada, argila, inconsolidada*	gestao : ***gerente***, *endomarketing, equipe,* ***gestor***, *analista*	porosimetro : *saturacao,* ***porosidade***, *compressibilidade, permeabilidade, densidade*
hornblenda : ***plagioclasio***, *augita, plagioclasios,* ***feldspatos***, *epidoto*	oceanografia : *doris, professor, mestrando, suplente, geologa*	densimetro : ***densidade***, *viscosidade, velocidade, massa, temperatura*
anidrita : *precipitacao,* ***calcio***, *esmectita, ilita, magnesianas*	enfermagem : *executante, bibliotecario, avaliador, equipe,* ***enfermeiro***	multimetro : ***voltagem***, *velocidade,* ***tensao***, ***corrente***, *frequencia*

Por fim, é relevante ressaltar que as avaliações baseadas em analogias semânticas possuem algumas limitações, por assumirem regularidades linguísticas que podem não ser integralmente mantidas em cenários reais de uso, uma vez que a semântica inerente à linguagem natural é intrinsecamente mais complexa do que se pode representar por simples operações lineares assumidas pelas analogias (LINZEN, 2016; ROGERS *et al.*, 2017; NEWMAN-GRIFFIS *et al.*, 2017). Portanto, acreditamos que as analogias de palavras podem fornecer importantes insumos para uma melhor compreensão sobre o espaço semântico dos modelos vetoriais, devendo, entretanto, ser combinadas com outras avaliações complementares para fornecer evidências adequadas para uma avaliação consistente. Nesse sentido, uma das principais motivações para este trabalho se refere justamente a oferecer uma ampla cobertura para tais avaliações, descritas ao longo do Capítulo IV.

4.4.2 Coerência de espaço semântico

Em função da reconhecida escassez de *benchmarks* e dados anotados para o domínio de O&G em português, conforme abordado ao longo desta tese, optamos por conduzir a tarefa de avaliação por *coerência de espaço semântico* segundo uma abordagem qualitativa, com o objetivo de explorar algumas das propriedades geométricas do espaço semântico dos modelos **`PetroVec`**, ao invés de estimar métricas quantitativas conforme reportado em trabalhos anteriores, tradicionalmente baseadas em métodos tais como *intruder word* (SCHNABEL *et al.*, 2015) ou *outlier detection* (CAMACHO-COLLADOS e PILEHVAR, 2018).

Conforme relatado por SCHNABEL *et al*. (2015) e GLADKOVA e DROZD (2016), um espaço semântico adequado deve ser organizado de maneira a fornecer regiões de vizinhança coerentes para cada vetor de palavras. Ou seja, o modelo de *word embedding* deve ser capaz de prover agrupamentos de termos semanticamente similares (CAMACHO-COLLADOS e PILEHVAR, 2018). Portanto, seguindo uma abordagem apresentada em trabalhos anteriores, conforme reportado por TSHITOYAN *et al*. (2019), WANG *et al*. (2018) e CAMACHO-COLLADOS e PILEHVAR (2018), realizamos uma primeira análise exploratória do espaço vetorial do modelo **`PetroVec-O&G`**, visualizado em uma projeção bidimensional utilizando o algoritmo t-SNE (VAN DER MAATEN e HINTON, 2008) para redução de dimensionalidade de 100 para 2 dimensões. O objetivo

dessa análise foi fornecer uma intuição sobre a organização do espaço semântico, avaliando a formação de grupos de palavras (*clusters*) mutuamente relacionadas. A Figura 4.6 ilustra uma projeção bidimensional t-SNE contemplando os 20 mil termos mais frequentes do *corpus especializado* de domínio. A área em destaque apresenta a região de vizinhança para o termo *'jurassico'*, que corresponde a um período geológico. É possível observar que os termos vizinhos mais próximos estão todos relacionados ao mesmo conceito (período geológico), sugerindo que o modelo foi capaz de corretamente capturar as propriedades semânticas que caracterizam esses termos, atribuindo seus vetores a posições vizinhas com significativa coesão. Uma versão completa e interativa desse gráfico está disponível no GitHub do **`PetroVec`**[51], implementado em Python usando a biblioteca Bokeh[52], que permite ao usuário explorar uma visão geral da estrutura do espaço semântico e ter uma intuição das similaridades entre as palavras. Adicionalmente, implementamos no site ***Petrolês*** uma ferramenta avançada para permitir experimentações e análises exploratórias mais detalhadas dos modelos **`PetroVec`**: o ambiente interativo do ***Visualizador de Espaço Semântico dos modelos PetroVec***[53]. A ferramenta foi implementada em Python com base no projeto de código aberto do Google: *Tensorflow Embedding Projector* (SMILKOV *et al.*, 2016), e permite experimentar diversas operações de similaridade entre termos, projeções do espaço semântico em 2D ou 3D usando PCA ou t-SNE, selecionar termos específicos, interagir com o gráfico e analisar regiões de vizinhança, conforme exemplo ilustrado pela Figura 4.7.

[51] Códigos utilizados na avaliação dos modelos PetroVec, disponíveis em: https://github.com/Petroles/Petrovec/tree/master/source/evaluation
[52] Bokeh: Python library for interactive visualization. https://bokeh.org/
[53] Visualizador de Espaço Semântico dos modelos PetroVec: http://petroles.ica.ele.puc-rio.br/projector.html

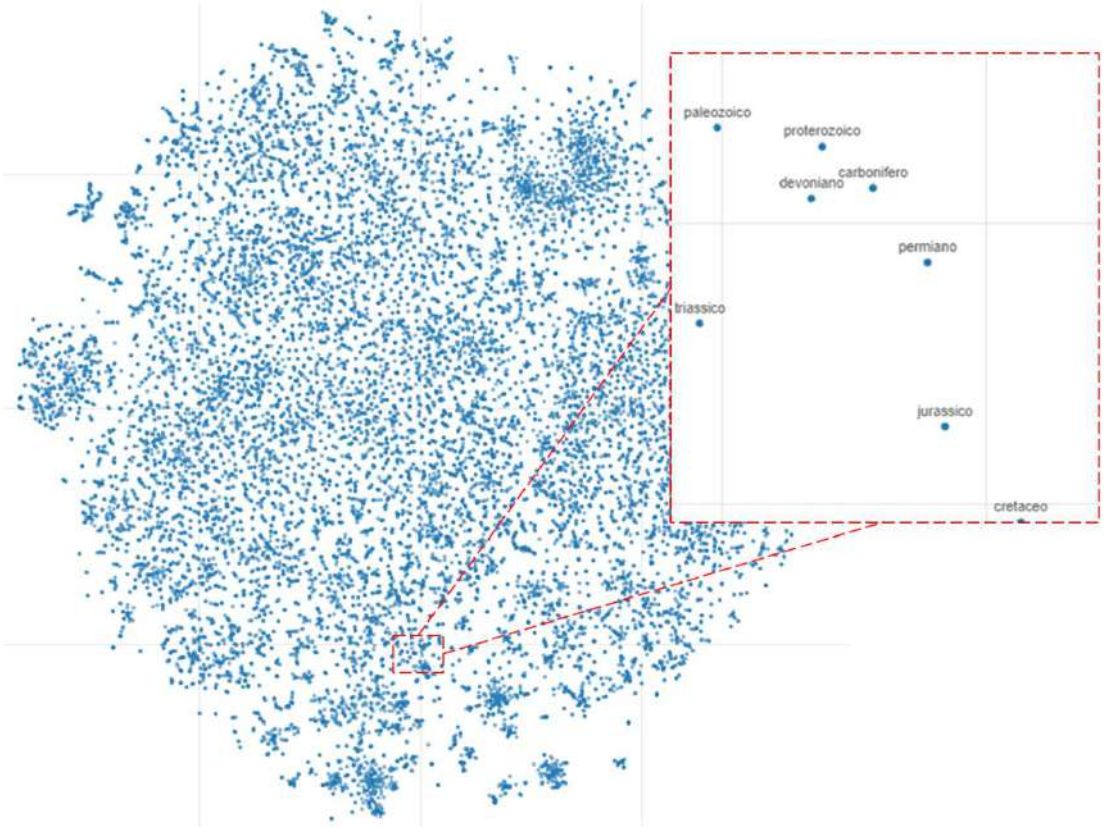


Figura 4.6 Projeção t-SNE bidimensional do espaço semântico para as 20.000 palavras mais frequentes do corpus, com destaque para região de vizinhança do termo *'jurassico'*.
Fonte: GOMES *et al.* (2021)

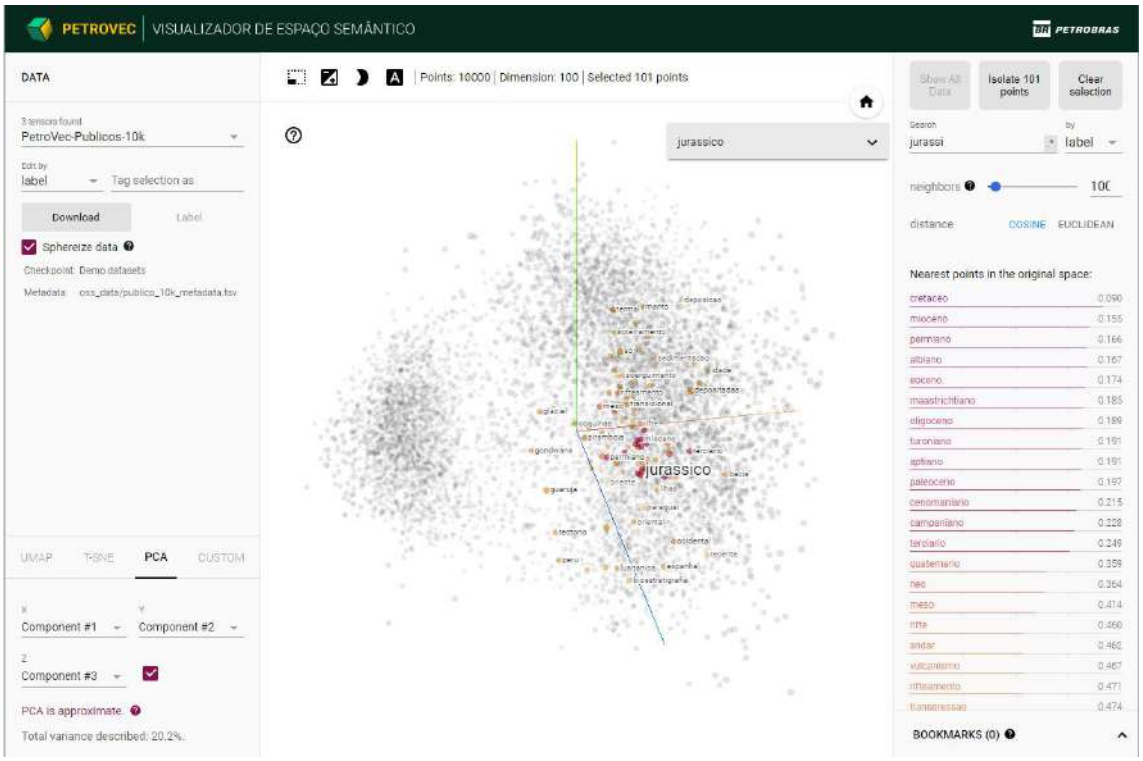


Figura 4.7 Ambiente de Visualiação do Espaço Semântico para os modelos PetroVec. Neste exemplo, foi selecionado um termo ('*jurassico*'), permitindo visualizar sua região de vizinhança em uma projeção 3D usando PCA, e os respectivos índices de similaridade dos termos mais próximos.

Em seguida, realizamos uma análise mais detalhada sobre a coerência do espaço semântico, avaliando se o modelo é capaz de manter a coesão perante diferentes grupos de palavras semanticamente relacionadas. Para essa tarefa, utilizamos como referência o *Schlumberger Oilfield Glossary*[54] (SOG), um reconhecido glossário em inglês para a área de O&G, e selecionamos quatro subdomínios relevantes, representando conceitos das áreas de: *(i)* *GEOLOGIA*, *(ii)* *PERFURAÇÃO*, *(iii)* *GEOFÍSICA*, e *(iv)* *DERIVADOS DE PETRÓLEO*. Então, para cada subdomínio, selecionamos 6 termos a partir do SOG, conferindo junto ao Dicionário do Petróleo para obter sua correta representação em português. A Tabela 4.12 detalha os quatro subdomínios e seus correspondentes termos utilizados para esta tarefa. O modelo **`PetroVec-O&G`** foi então utilizado para calcular a distância cosseno entre cada par de palavras, criando assim uma matriz de índices de similaridade.

Para analisar os resultados dessa matriz, duas visualizações alternativas para esses dados são propostas: *(i)* como um gráfico de mapa de calor (*heatmap*), destacando suas métricas de similaridade ilustradas na forma de gradientes de cor (Figura 4.8); e *(ii)* como um diagrama do tipo *chord*, em que as relações de similaridade são representadas como a intensidade das ligações ente os elementos (Figura 4.9). O gráfico de *heatmap (i)* fornece uma visão geral da similaridade entre os termos, enfatizando a formação de grupos coesos contendo elementos da mesma categoria concentrando altos índices de similaridade, enquanto elementos de categorias diferentes são exibidos em tonalidades amareladas que representam baixa similaridade. O gráfico *chord (ii)* destaca que a interconexão entre os termos é predominante entre os elementos do mesmo subdomínio, com poucas exceções conectando elementos de categorias diferentes. Ambos os gráficos reforçam a formação de grupos distintos e coesos, com mínima sobreposição de similaridade entre categorias diferentes, sugerindo que o modelo **`PetroVec`** foi capaz de codificar corretamente um espaço semântico fortemente coerente.

Tabela 4.12: Listagem de termos associados a cada subdomínio de O&G

Geologia	***Perfuração***	***Geofísica***	***Derivados de Petróleo***
anidrita	valvula	eletromagnetica	nafta
evaporito	duto	espectral	querosene
calcarenito	riser	amplitude	glp
argila	pig	reflexao	gasolina
arenito	choke	acustica	diesel
folhelho	anm	sismica	combustivel

[54] Schlumberger Oilfield Glossary: https://www.glossary.oilfield.slb.com

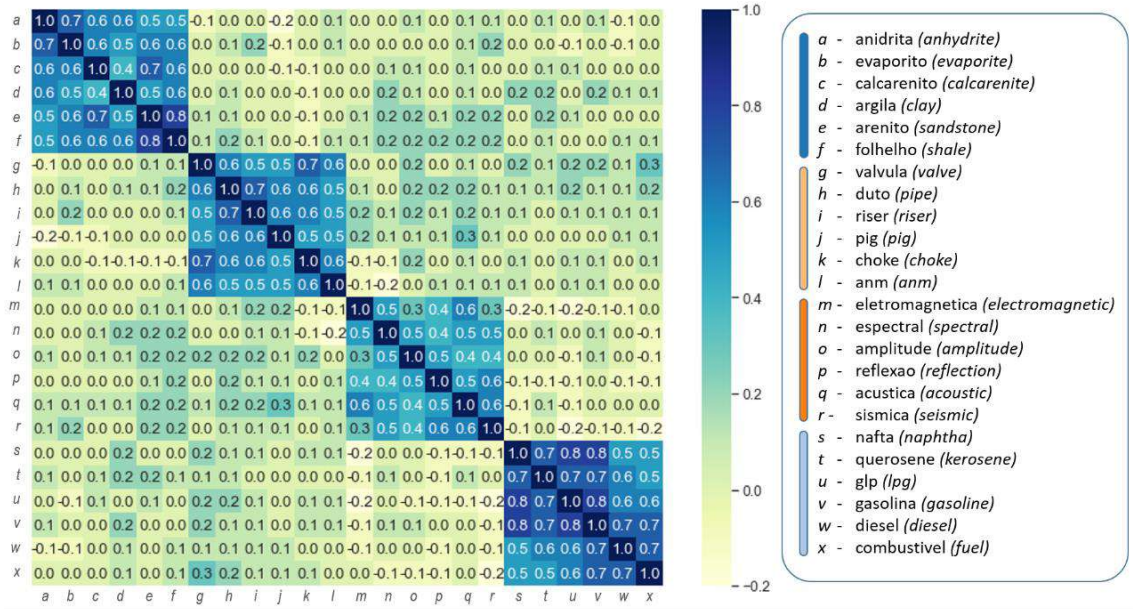


Figura 4.8 Gráfico *heatmap* para a matriz de similaridade, evidenciando a formação de grupos coesos dentro de cada subdomínio.
Fonte: GOMES *et al.* (2021)

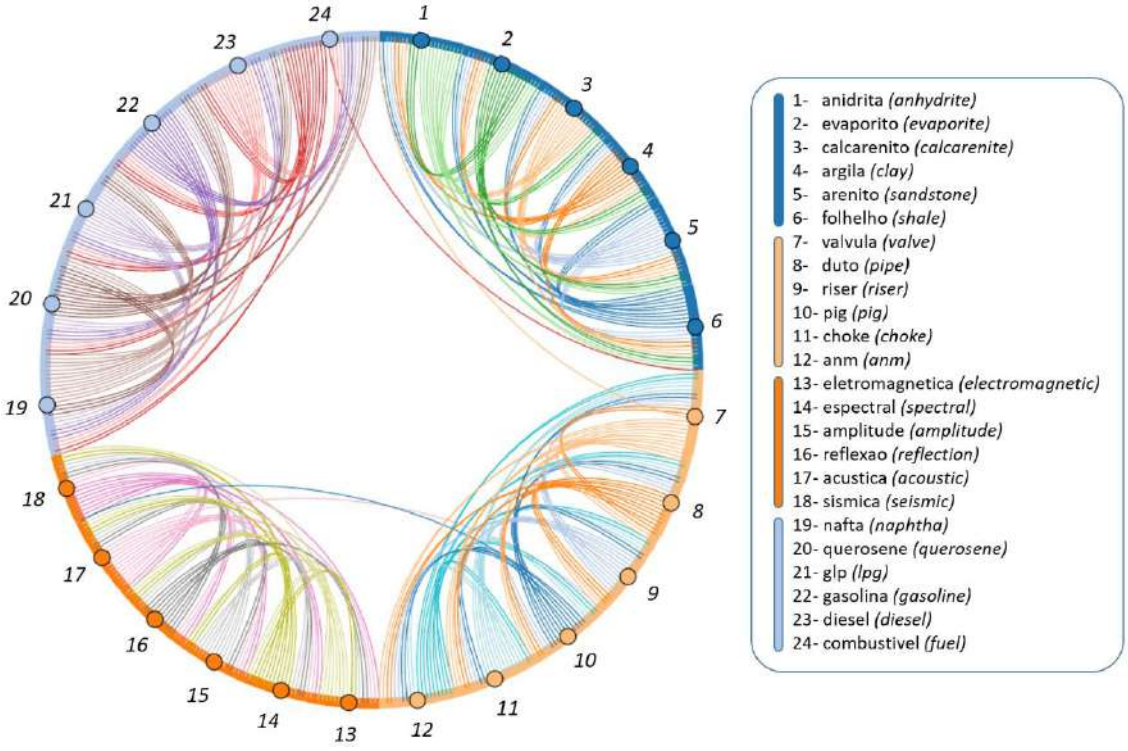


Figura 4.9 Gráfico *chord* para a matriz de similaridade, evidenciando a predominância de ligações fortes entre elementos de uma mesma categoria (*são exibidas relações com índices de similaridade maiores que 0,3, uma vez que índices menores que este valor são considerados irrelevantes*).
Fonte: GOMES *et al.* (2021)

4.4.3 Categorização de conceitos

Em complemento às avaliações de coerência de espaço semântico apresentadas na seção anterior, realizamos análises de categorização de conceitos (*concept categorization*) para avaliar a capacidade do modelo em prover agrupamentos de termos semanticamente relacionados no domínio, comparativamente validados por um algoritmo automático de clusterização (*clustering*). Seguindo uma metodologia reportada por BARONI *et al.* (2014), SCHNABEL *et al.* (2015), PADARIAN *et al.* (2019) e WANG *et al.* (2019), o método consiste em, dado um conjunto de termos semanticamente relacionados a categorias predefinidas, o resultado de um algoritmo de clusterização (como o *k*-means) deve ser capaz de corretamente identificar os agrupamentos dos termos similares de forma a corresponder às respectivas categorias originalmente pré-estabelecidas.

Portanto, para esta análise, selecionamos manualmente cinco termos de referência, semanticamente relacionados a diferentes categorias relevantes no domínio de O&G: *(i) carapeba* (campos de produção); *(ii) combustivel* (derivados de petróleo); *(iii) paleoceno* (período geológicos); *(iv) geofisica* (disciplina científica de interesse para o domínio); e *(v) arenito* (subdomínio de geologia). Em seguida, para cada um dos cinco termos de referência, utilizamos o modelo **`PetroVec`** para calcular o conjunto das dez palavras mais similares, obtendo as regiões de vizinhança para cada termo. Os 55 vetores resultantes foram empilhados em uma matriz, por sua vez submetida ao algoritmo de clusterização *k*-means. Por fim, os resultados foram projetados em um espaço bidimensional usando t-SNE e os agrupamentos resultantes do *k*-means sinalizados por diferentes cores, conforme ilustrado na Figura 4.10. É possível observar que o algoritmo de clusterização foi capaz de corretamente identificar os agrupamentos, associando cada um dos termos à sua respectiva categoria de referência, conforme esperado, sem sobreposição de elementos pertencentes a grupos diferentes, sugerindo uma significativa coesão desses conceitos em agrupamentos distintos.

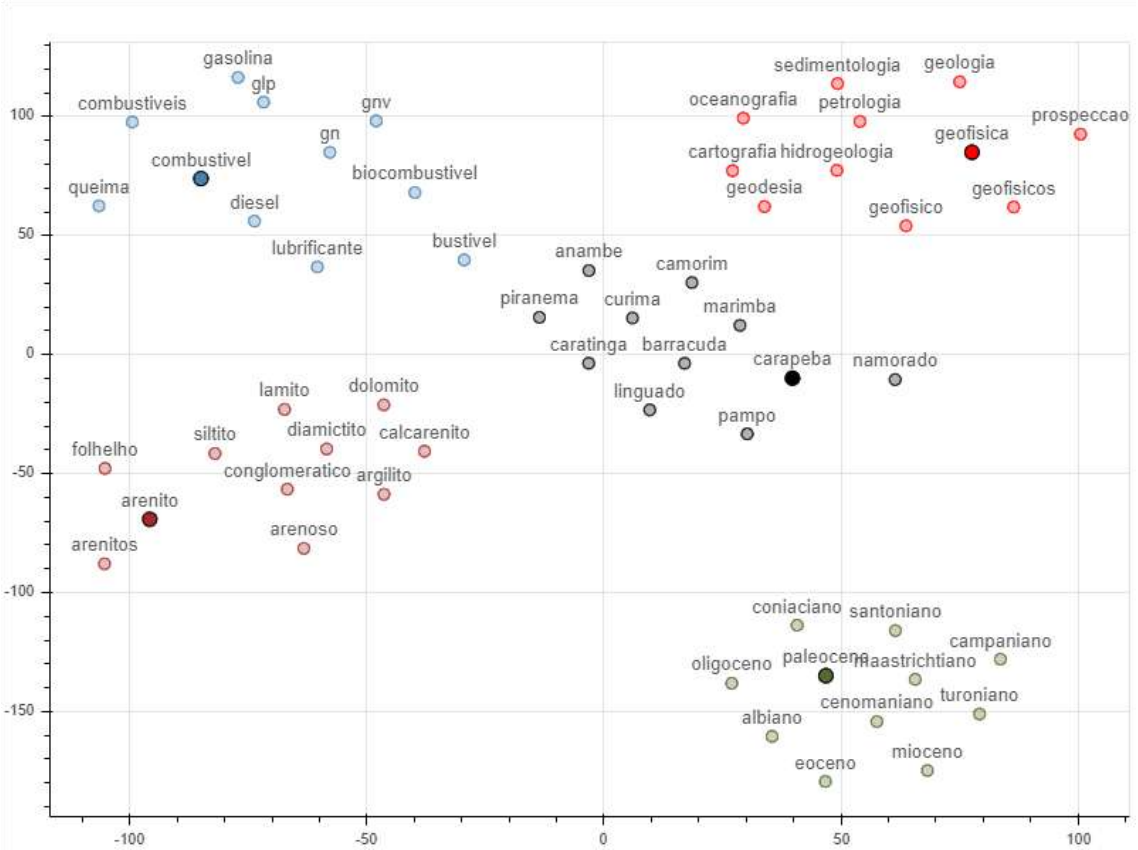


Figura 4.10 Projeção t-SNE com o resultado do algoritmo *k*-means para as regiões de vizinhança dos termos de referência (*sinalizados com destaque de cor em cada agrupamento*). O algoritmo identificou corretamente as categoriais originalmente predefinidas.
Fonte: GOMES *et al.* (2021)

4.4.4 Análises exploratórias complementares

Além das avaliações qualitativas previamente apresentadas, realizamos outros experimentos complementares focando em um conjunto de termos técnicos que reconhecidamente possuem particularidades semânticas que são específicas do domínio de O&G. Para essa análise, selecionamos um conjunto amostral de termos técnicos do domínio e utilizamos os modelos **`PetroVec-O&G`** e **`skipgram-NILC`** para listar suas respectivas *top*-3 palavras mais similares. Os resultados obtidos em cada modelo são comparativamente apresentados na Tabela 4.13. É possível observar que as palavras retornadas pelo modelo especializado estão todas semanticamente relacionadas ao termo de referência considerando o contexto de O&G, sugerindo a capacidade do modelo **`PetroVec-O&G`** de capturar propriedades semânticas a partir do corpus de domínio. Em contrapartida, o modelo **`skipgram-NILC`** retorna palavras cujo significado denota conceitos genéricos e com semântica não relacionada ao domínio de O&G, predominantemente influenciado pela natureza mais abrangente da composição de seu corpus de treinamento.

Tabela 4.13: Resposta dos modelos para termos técnicos da área de O&G

Termo	`PetroVec-O&G`	`skipgram-NILC`
campo	reservatorio poco dominio	estádio gramado farião
rocha	porosidade geradora porosa	ribeiro alves moreira
falha	falhamento alinhamento descontinuidade	pane avaria desatenção
duto	oleoduto tubulacao cilindro	tubo aquecedor orifício
óleo	petroleo vazar liquido	sabão alcatrão alumínio
choke	succao valvula bop	defying remorse disciple
árvore	anm semisubmersivel submarina	planta trepadeira frondosa
fadiga	flexao desgaste tracao	irritabilidade sonolencia salivacao
calado	costado navio conves	silvino revez quieto
bcs	centrifugo equipado bombeamento	cfd nfs nist

Uma das áreas com características de nomenclatura mais peculiares no domínio de O&G corresponde às denominações associadas aos campos de petróleo, que comumente referenciam nomes de lugares ou animais marinhos. A correta representação desses termos é de fundamental importância para o adequado funcionamento dos algoritmos de PLN aplicados a situações reais de uso, que são diretamente dependentes da capacidade de generalização do modelo vetorial. Nesta etapa do experimento, selecionamos um conjunto de termos associados a campos de petróleo, obtidos a partir da listagem completa dos campos regulamentados pela ANP [55] no Brasil, conforme reproduzido no Apêndice A.2. Para cada termo, listamos seus respectivos elementos mais

[55] Planos de Desenvolvimento dos campos, ANP. Disponível em: http://www.anp.gov.br/exploracao-e-producao-de-oleo-e-gas/gestao-de-contratos-de-e-p/fase-de-producao/planos-de-desenvolvimento

similares conforme retornados pelos modelos **PetroVec-O&G** e **skipgram-NILC**, cujos resultados são apresentados na Tabela 4.14. Primeiramente, é importante destacar a ausência de alguns termos no vocabulário do modelo **skipgram-NILC**, o que poderia prejudicar sua aplicação em um cenário real de uso voltado para a indústria de O&G. Para os demais casos, é possível notar uma ampla divergência de significado assumido pelo modelo genérico, notadamente influenciado pela semântica generalista presente em seus corpora de contexto geral. O modelo especializado **PetroVec-O&G**, por sua vez, identifica com relativo sucesso as relações de similaridade, corretamente associando-os a outros nomes de campos de petróleo. Em tempo, cabe ressaltar que os modelos não foram explicitamente alimentados com essa relação de nomes de campos, que foi utilizada apenas para os testes qualitativos. As relações de similaridade aqui reportadas foram, portanto, automaticamente capturadas pelo modelo de forma não-supervisionada a partir do *corpus especializado* de O&G criado nesta pesquisa.

Tabela 4.14: Resposta dos modelos para nomes de campos de petróleo

Termo	PetroVec-O&G	skipgram-NILC
lula	sapinhoa jubarte bauna	ex-presidente inácio inpcio
acaua	angico tabuiaiá seriema	-
albacora	roncador pampo jubarte	patudo espadarte lagostim
aratu	buracica jiquia cassarogongo	itapoan jangadeiro alterosa
carapanauba	araracanga abalone pajeu	-
carapeba	pargo linguado cherne	órleans ofmcap anaia
cherne	linguado anequim badejo	entrecosto gambas gratinada
enchova	badejo pampo corvina	savelha colhereiros curimbatá
espada	xareu curima atum	adaga armadura katana

furado	caioba robalo guaricema	desgovernado caminhão carregado
jubarte	marimba pirauna cachalote	cachalote baleia megapode
linguado	badejo corvina cherne	badejo arenque bacalhau
malhado	congro garoupinha corvina	brilhoso atarracado esbelto
namorado	pampo marimba cherne	ex-namorado ex-marido marido
pirauna	parati bicudo polvo	-
ubarana	macau taquipe pescada	ipanguaçu severínia pontalinda
viola	voador pirauna parati	percussão trompete sanfona
voador	cachalote parati polvo	camaleão pássaro chocalho
xareu	espada curima atum	paleosuchus lince-do-deserto attiéké

Capítulo V

Conclusões e Trabalhos Futuros

"In literature and in life we ultimately pursue, not conclusions, but beginnings.."

Sam Tanenhaus

Este capítulo apresenta um resumo sobre o trabalho de pesquisa realizado nesta tese. Os principais resultados são discutidos e analisados quanto ao atendimento da hipótese central e dos objetivos propostos por esta tese. Por fim, analisamos perspectivas para a evolução da pesquisa como tema de trabalhos futuros.

5.1 Conclusões

Neste trabalho, apresentamos o **PetroVec**, um conjunto de modelos de vetorização de palavras em português especializados no domínio de Óleo e Gás. Para viabilizar o treinamento dos modelos, criamos um corpus específico do domínio, composto por milhares de documentos técnicos e científicos publicados em português por Universidades e principais instituições de referência na indústria nacional de O&G. O corpus especializado de domínio contempla mais de 85 milhões de tokens e é atualmente o maior conjunto textual público reportado para o domínio de O&G neste idioma. Os modelos vetoriais **PetroVec** foram treinados em dois dos principais métodos disponíveis (Word2Vec e FastText) e representam um dos principais fundamentos para o desenvolvimento de algoritmos de PLN aplicados a cenários reais acadêmicos e industriais.

A fim de oferecer evidências para uma avaliação consistente da qualidade das representações geradas pelos modelos **PetroVec**, realizamos uma detalhada e abrangente cobertura de testes e avaliações, contemplando metodologias quantitativas baseadas em análises intrínseca e extrínseca, além de um conjunto de avaliações qualitativas para explorar propriedades linguísticas codificadas no espaço semântico dos modelos. Adicionalmente, comparamos resultados obtidos pelos modelos **PetroVec** em relação a um modelo público de referência, treinado a partir de um corpus de contexto geral em português (HARTMANN *et al.*, 2017), além de um modelo especializado de O&G desenvolvido em um trabalho preliminar do autor desta tese (GOMES *et al.*, 2018), que serviram como *baseline* para as métricas.

Para viabilizar a realização das metodologias de avaliação *intrínseca* e *extrínseca*, conduzimos uma iniciativa em parceria com pesquisadores das Universidades UFRGS e PUC-RS, apoiada pelo Centro de Pesquisa e Desenvolvimento da Petrobras (Cenpes). Para a ***avaliação intrínseca*** (apresentada na Seção §4.2), criamos um *dataset* composto por 1500 pares de termos relevantes para o domínio e seus respectivos índices de similaridade semântica, anotados por especialistas da Petrobras e da comunidade acadêmica, o que permitiu o desenvolvimento de uma métrica para os modelos. Os resultados demonstram que os modelos especializados **PetroVec** obtiveram desempenho superior ao modelo genérico em todos os critérios, principalmente quanto à cobertura de vocabulário. Nesses testes, os modelos treinados com o algoritmo

Word2Vec apresentaram resultado ligeiramente melhor que o FastText. Em relação à ***avaliação extrínseca*** (§4.3), realizamos uma metodologia para analisar a qualidade dos modelos quando aplicados a uma tarefa específica de reconhecimento de entidades nomeadas (REN) no domínio de Geociências, utilizando uma versão revisada e estendida do GeoCorpus (AMARAL, 2017). Nessa análise, os modelos especializados **`PetroVec-O&G`** e **`PetroVec-hybrid`** igualmente apresentaram resultados consistentemente superiores aos modelos genérico e *baseline*, obtidos principalmente em função de sua maior revocação. Contrariamente ao observado na análise intrínseca, neste teste os melhores resultados foram obtidos pelo algoritmo FastText, alcançando incrementos de revocação de até 8 pontos percentuais em relação às suas contrapartes Word2Vec, sugerindo que a propriedade de lidar com informações de subpalavras do FastText foi determinante para esses resultados. Por fim, para as duas metodologias (intrínseca e extrínseca), ampliamos as análises fornecendo métricas de significância estatística para confirmar os resultados obtidos pelo **`PetroVec`** em comparação aos demais modelos de referência.

As *análises qualitativas*, por sua vez, concentraram-se em avaliar o modelo **`PetroVec-O&G`**, treinado exclusivamente a partir do *corpus especializado* de domínio, e contemplaram abordagens baseadas nas técnicas de *analogias de palavras*, *coerência de espaço semântico* e *categorização de conceitos* a partir do espaço vetorial (TURIAN et al., 2010; SCHNABEL et al., 2015; MUNEEB et al., 2015). Observamos resultados promissores com operações de ***analogias de palavras*** (§4.4.1), reproduzindo alguns dos comportamentos originalmente relatados por MIKOLOV *et al.* (2013a), inclusive encontrando sucesso em interessantes cenários envolvendo conceitos de representações bilíngues. As analogias foram investigadas em maior profundidade através de um método estatístico para refutar a hipótese de que poderiam ser um artefato de corpus. Além disso, aplicamos o método de *analogia reversa* (LINZEN, 2016) para ratificar que a qualidade dos resultados foi mantida ao inverter a direção das operações. Em relação à ***coerência de espaço semântico*** (§4.4.2), realizamos análises exploratórias de projeções bidimensionais do espaço vetorial usando t-SNE para obter uma visão geral sobre a formação de agrupamentos de termos similares. Além disso, calculamos a matriz de similaridade entre um conjunto de termos previamente selecionados, fornecendo visualizações para interpretação dos dados na forma de gráficos *heatmap* e *chord*. Os resultados evidenciaram a formação de agrupamentos coesos entre os termos e

confirmaram um espaço semântico fortemente coerente. A partir do método de ***categorização de conceitos*** (§4.4.3), observamos no espaço semântico do modelo a formação de agrupamentos coesos com termos semanticamente relacionados no domínio, confirmados a partir do algoritmo de clusterização *k*-means. Por fim, a partir de ***análises exploratórias complementares*** (§4.4.4), analisamos as regiões de vizinhança para amostras do vocabulário técnico de O&G que apresentam particularidade semânticas específicas do domínio. Os resultados sugerem uma qualidade superior obtida pelos modelos especializados em comparação ao modelo genérico para esses conjuntos amostrais, apresentando melhor cobertura de vocabulário de domínio e regiões de similaridade semanticamente mais significativas.

Portanto, os resultados de todas as avaliações convergem ao evidenciar que os modelos especializados **`PetroVec`** superaram o modelo genérico em testes focados no domínio de O&G, sugerindo que os modelos foram capazes de automaticamente codificar propriedades sintáticas e semânticas específicas do vocabulário técnico de forma não-supervisionada a partir do corpus especializado de treinamento. Além disso, é relevante destacar que, mesmo diante de uma menor disponibilidade de dados, a especificidade do corpus no domínio representou um aspecto crucial para o sucesso do treinamento, ratificando as conclusões reportadas por trabalhos anteriores (LAI *et al.*, 2016; PAKHOMOV *et al.*, 2016; NOORALAHZADEH *et al.*, 2018; WANG *et al.*, 2018; ALSENTZER *et al.*, 2019; TSHITOYAN *et al.*, 2019; PADARIAN e FUENTES, 2019). Isto é, os modelos especializados **`PetroVec`** apresentaram resultados consistentemente superiores ao modelo genérico mesmo tendo sido treinados em um corpus significativamente menor (*i.e.*, o modelo genérico foi treinado em um corpus cerca de 16 vezes maior do que o *corpus específico de O&G* do **`PetroVec`**, conforme ilustrado na Figura 3.3). Ainda assim, de forma complementar ao domínio, o tamanho do corpus também desempenha papel importante na qualidade dos modelos, conforme pôde ser observado pelos resultados superiores obtidos pelo **`PetroVec`** em relação ao modelo **`baseline-O&G`** (este último, por sua vez, treinado em uma versão preliminar cerca de 8 vezes menor do *corpus específico de O&G*).

Dessa forma, perante os resultados apresentados nesta tese, suportados por uma consistente metodologia de avaliação dos modelos, é possível afirmar que a hipótese central (*H.1*) proposta para esta pesquisa foi plenamente validada:

(H.1) *Modelos de vetorização de palavras especializados no domínio de Óleo e Gás em português, treinados a partir de corpora específicos, são capazes de melhorar a qualidade de suas representações semânticas e o seu desempenho quando aplicados em tarefas de PLN específicas do domínio, comparativamente em relação a um modelo pré-treinado a partir de corpora de contexto genérico?*

Em tempo, cabe destacar a enorme transformação pela qual vem passando a área de pesquisa em PLN nos últimos dois anos, especialmente motivada pelo advento de uma nova geração de modelos de representação contextual, baseados em complexas arquiteturas de aprendizagem profunda (conforme apresentado em §2.1.3). Esses modelos alcançaram impressionantes resultados e estabeleceram novos patamares de estado-da-arte para diversas tarefas aplicadas de PLN. Entretanto, apesar dos resultados promissores, essas arquiteturas demandam um expressivo aumento nos custos computacionais necessários para o treinamento e inferência, muitas vezes inviabilizando sua utilização por projetos de pesquisa acadêmicos, ou sua efetiva implantação em produção. Além disso, estudos recentes demonstraram que, em muitos cenários envolvendo ambientes de produção com dados reais em escala industrial, os modelos estáticos obtiveram resultados altamente competitivos, oferecendo uma alternativa eficiente a uma fração do custo dos modelos contextuais mais complexos (conforme explorado em detalhes na Seção §2.1.5).

Portanto, ratificando a relevância e originalidade do tema de pesquisa (conforme levantamento de literatura descrito em §2.2 e §2.2.1), os modelos estáticos tais como o **`PetroVec`** são mais leves e rápidos para treinamento e inferência, tendo sido amplamente utilizados com sucesso em diversas aplicações reais no ambiente acadêmico, industrial e comercial. Atualmente, os modelos **`PetroVec`** encontram-se implantados no ambiente de produção da Petrobras (KRUEL *et al.*, 2019), dando suporte a um sistema de expansão automática de consulta para projetos relacionados a busca semântica.

Por fim, todos os recursos desenvolvidos no contexto desta tese estão publicamente disponíveis para a comunidade científica no repositório ***Petrolês*** [56], incluindo os corpora e modelos vetoriais pré-treinados **`PetroVec`**, além dos conjuntos de dados anotados para validação e os scripts para pré-processamento, treinamento e avaliação (disponíveis no repositório GitHub[57]). Além disso, no portal Petrolês, também disponibilizamos o *Visualizador de Espaço Semântico dos modelos PetroVec*[58], uma ferramenta interativa para testes e análise exploratória do espaço semântico dos modelos, em projeções 2D e 3D usando projeções t-SNE e PCA, além de permitir avaliar operações de similaridade entre termos do vocabulário técnico e explorar suas regiões de vizinhança.

Objetiva-se que o compartilhamento público deste material com a comunidade científica possa contribuir com o avanço da linha de pesquisa de PLN especializada no domínio de O&G, cuja literatura encontra-se carente de tais insumos. Acreditamos que diversos pesquisadores atuantes neste domínio, tanto na indústria como na Universidade, possam se beneficiar dos resultados deste trabalho. Alguns de aplicações práticas de PLN de interesse para o setor de O&G, incluem: *sistemas de segurança e predição de risco industriais* (BIRNIE *et al.*, 2019; CAI *et al.*, 2019), *extração da informação* (BLINSTON e BLONDELLE, 2017; WANG *et al.*, 2018), *otimização de operações de perfuração de poços* (WILSON, 2017; FURTADO, 2017; CASTIÑEIRA *et al.*, 2018; COLOMBO *et al.*, 2019, UCHEREK *et al.*, 2020), *sumarização de documentação técnica* (MARQUES *et al.*, 2019), *sistemas de perguntas e resposta* (JACOBS, 2019), e *classificação automática de texto* (NOORALAHZADEH *et al.*, 2018; KHABIRI *et al.*, 2019; SANCHEZ-PI *et al.*, 2014; RIBEIRO *et al.*, 2020).

5.2 Trabalhos Futuros

Considerando o cenário de aplicações de PLN especializadas no domínio de O&G no idioma português, este trabalho pode ser considerado pioneiro ao disponibilizar um conjunto de insumos que podem servir de base para fomentar o desenvolvimento de novas soluções e linhas de pesquisa nesse domínio. Portanto, diversas oportunidades podem ser vislumbradas como evolução deste trabalho em projetos futuros.

[56] Petrolês – Repositório de artefatos de PLN para o domínio de O&G: http://petroles.ica.ele.puc-rio.br/
[57] Repositório de código-fonte do PetroVec: https://github.com/Petroles/Petrovec
[58] Visualizador de Espaço Semântico dos modelos PetroVec: http://petroles.ica.ele.puc-rio.br/projector.html

Primeiramente, acreditamos haver espaço para melhorias na qualidade dos modelos **`PetroVec`** a partir da ampliação dos corpora de treinamento, e incorporando novas fontes de dados como publicações científicas, artigos, teses e dissertações, devidamente relacionadas aos temas de interesse do domínio. Além disso, é desejável a aquisição de bases curadas e recursos de conhecimento (*knowledge resources*) especializados no domínio, como ontologias e tesauros, de maneira a melhorar a qualidade de representação para elementos lexicais menos frequentes no corpus (FARUQUI *et al.*, 2015; PILEHVAR e COLLIER, 2016).

Adicionalmente, considerando-se a característica do vocabulário de O&G de comumente conter conceitos na forma de expressões multi-palavras (*Multi-Word Expressions*, MWE), é relevante considerar uma possível evolução dos modelos de forma a contemplar MWEs, especialmente considerando-se a eventual disponibilidade de bases de conhecimento anotadas e estruturadas (NEWMAN-GRIFFIS *et al.*, 2018). Nós conduzimos alguns experimentos preliminares nesse sentido, seguindo MIKOLOV *et al.* (2013a) e extraindo conceitos utilizando *Pointwise Mutual Information* (PMI) (PILEHVAR e CAMACHO-COLLADOS, 2020). Entretanto, após uma análise manual dos resultados obtidos a partir desse método automático, observamos que as MWEs não se mostraram semanticamente coerentes para serem utilizadas. Isto é, a maioria dos termos encontrados, apesar dos seus altos *scores* PMI, de fato não carregavam nenhum significado linguístico relevante para serem consideradas expressões válidas. Portanto, conforme descrito por CONSTANT *et al.* (2017) e NEWMAN-GRIFFIS *et al.* (2018) seria ideal dispor de bases de conhecimento estruturadas e manualmente curadas para viabilizar esta funcionalidade. Portanto, em função da atual escassez de bases de referência e da pouca qualidade obtida a partir de métodos automáticos, acreditamos que a incorporação de MWEs demanda pesquisa e desenvolvimento adicionais a serem abordados em trabalhos futuros.

No decorrer do tratamento do corpus, observamos a ocorrência de muitos ruídos, potencialmente relacionados a erros de grafia ocasionados por problemas de extração textual a partir dos documentos em PDF. Considerando que a qualidade dos modelos é diretamente afetada por esses ruídos no corpus de treinamento, é desejável dispor de melhores técnicas de extração de texto a partir dos PDFs, sejam técnicas mais avançadas para reconhecimento de caracteres (*Optical Character Recognition*, OCR), novas estratégias de pré-processamento para filtrar e eliminar trechos ruidosos, ou pelo

desenvolvimento de métodos mais avançados para correção gramatical através de dicionários, modelos de linguagem ou outros algoritmos automatizados mais sofisticados.

Considerando as recentes técnicas para representações contextuais como ELMO, BERT e GPT, uma evolução natural deste trabalho consiste em explorar o desenvolvimento eficiente dessas abordagens para o domínio de O&G em português, utilizando variações de composição dos corpora especializados e corpora híbridos. Alguns trabalhos recentes foram propostos com o objetivo de oferecer versões especializadas de modelos contextuais para domínios específicos, em especial na área biomédica para o idioma inglês (BELTAGY *et al.*, 2019; ALSENTZER *et al.*, 2019; LEE *et al.*, 2020).

Além disso, é desejável estabelecer novas soluções que permitam avaliar os modelos utilizando metodologias extrínsecas, obtendo uma maior representatividade do espectro de aplicações em que os modelos vetoriais podem ser utilizados em problemas de PLN, como busca semântica, classificação automática e similaridade de documentos.

Por fim, cabe mencionar que esta pesquisa faz parte de uma iniciativa interinstitucional mais ampla, liderada pelo Centro de Pesquisas da Petrobras em parceria com algumas das principais Universidades brasileiras, envolvendo um grupo de pesquisadores com o objetivo de desenvolver novas tecnologias para avançar o estado-da-arte em soluções de PLN aplicadas ao domínio de O&G em português. Portanto, algumas das sugestões apresentadas como trabalhos futuros encontram-se atualmente previstas no contexto dessa iniciativa. A Figura 5.1 ilustra um panorama simplificado com algumas das pesquisas em andamento e as principais instituições de ciência e tecnologia envolvidas.

Figura 5.1 Diagrama ilustrativo de algumas das linhas de pesquisa conduzidas por uma iniciativa interinstitucional para o deselvolvimento de soluções de PLN para o setor de O&G em português.

Referências Bibliográficas

ACL. **What's new, different and challenging in ACL 2019?**, 2019. Disponível em: <https://acl2019pcblog.fileli.unipi.it/?p=156>. Acesso em: 13 jan. 2021

AGIRRE, E., ALFONSECA, E., HALL, K., *et al*. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: **Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. Boulder, Colorado: Association for Computational Linguistics, 2009, p. 19–27. Disponível em: <https://www.aclweb.org/anthology/N09-1003>. Acesso em: 15 jan. 2021.

AKBIK, A., BLYTHE, D., VOLLGRAF, R. Contextual String Embeddings for Sequence Labeling. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, p. 1638–1649, 2018.

ALLAHYARI, M.; POURIYEH, S. A.; ASSEFI, M.; SAFAEI, S., TRIPPE, E., GUTIERREZ, J., KOCHUT, K. **Text Summarization Techniques: A Brief Survey**. International Journal of Advanced Computer Science and Applications (IJACSA). 8. 397-405. 10.14569/IJACSA.2017.081052, 2017.

ALSENTZER, E., MURPHY, J., BOAG, W., WENG, W.-H., JINDI, D., NAUMANN, T., MCDERMOTT, M. Publicly Available Clinical BERT Embeddings, **in: Proceedings of the 2nd Clinical Natural Language Processing Workshop**, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 72-78, 2019.

AMARAL, D. **Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras**, D.Sc. thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2017. Disponível em: <http://tede2.pucrs.br/tede2/handle/tede/8035>. Acesso em 10/01/2021

ARORA, S., MAY, A., ZHANG, J., RÉ, C. Contextual embeddings: When are they worth it?, **in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, Association for Computational Linguistics, Online, pp. 2650-2663. doi: 10.18653/v1/2020.acl-main.236, 2020,

BAHDANAU, D., CHO, K., BENGIO, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. **CoRR**, abs/1409.0473.

BAKAROV, A. **A Survey of Word Embeddings Evaluation Methods**. arXiv:1801.09536 [cs], 2018. Disponível em: <http://arxiv.org/abs/1801.09536>. Acesso em: 15 jan. 2021.

BARONI, M., DINU, G., KRUSZEWSKI, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. **In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**

(Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, pp. 238-247, 2014.

BAST, H., BUCHHOLD, B., HAUSSMANN, E. (2016). **Semantic Search on Text and Knowledge Bases**. Foundations and Trends® in Information Retrieval. 10. 119-271. doi:10.1561/1500000032.

BELTAGY, I., LO, K., COHAN, A. SciBERT: A Pretrained Language Model for Scientific Text. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Association for Computational Linguistics, pp. 3615–3620, 2019.

BENAICH, N., HOGARTH, I. **State of AI Report 2020**. Disponível em: <https://www.stateof.ai/>. Acesso em: 12 jan. 2021

BENDER, E., GEBRU, T., MCMILLAN-MAJOR, A., SHMITCHELL, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: **Conference on Fairness, Accountability, and Transparency (FAccT '21)**, March, 2021, Virtual Event, Canada. ACM. doi: 10.1145/3442188.3445922, 2021.

BENGIO, Y., SIMARD, P., FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. **IEEE Trans. Neural Networks 5**, 157–166 (1994).

BENGIO, Y., DUCHARME, R., VINCENT, P., JANVIN, C. (2003). A Neural Probabilistic Language Model. **The Journal of Machine Learning Research**, 3, 1137-1155, 2003.

BIRNIE, C., SAMPSON, J., SJAASTAD, E., JOHANSEN, B., OBRESTAD, L., LARSEN, R., KHAMASSI, A. Improving the Quality and Efficiency of Operational Planning and Risk Management with ML and NLP, in: **SPE Offshore Europe Conference and Exhibition**, Society of Petroleum Engineers, Aberdeen, UK, 2019.

BLEI, D., NG, A., JORDAN, M., Latent dirichlet allocation. **Journal of machine Learning research**, vol. 3, no. Jan, pp. 993–1022, 2003

BLINSTON, K., BLONDELLE, H. **Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents**. The Leading Edge, 2017.

BOJANOWSKI, P., GRAVE, E., JOOULIN, A. MIKOLOV, T. Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**. 2017.

BRUNI, R., TRAN, N., BARONI, M. Multimodal distributional semantics, **Journal of Artificial Intelligence Research 49**, p. 1-47, 2014.

CAI, S., PALAZOGLU, A., ZHANG, L., HU, J. Process alarm prediction using deep learning and word embedding methods, **ISA Transactions 85**, 274-283, 2019.

CAMACHO-COLLADOS, J., PILEHVAR, M. From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. **Journal of Artificial Intelligence Research** 63, 2018.

CASTIÑEIRA, D., TORONYI, R., SALERI, N. **Machine Learning and Natural Language Processing for Automated Analysis of Drilling and Completion Data**, Society of Petroleum Engineers, 2018.

CER, D., YANG, Y., KONG, S.-Y., HUA, N., KIMTIACO, N., JOHN, R. S., CONSTANT, N., GUAJARDO-CESPEDES, M., YUAN, S., TAR, C., SUNG, Y.-H., STROPE, B., KURZWEIL, R. Universal sentence encoder. **in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, Association for Computational Linguistics. doi: 10.18653/v1/D18-2029, Brussels, Belgium, 2018.

CHO, K., MERRIENBOER, B. van, GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., BENGIO, Y. Learning phrase representations using rnn encoder–decoder forstatistical machine translation. **In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),**. [S.l.]: Association for Computational Linguistics, 2014. p. 1724–1734.

CLARK, K., LUONG, M., LE, Q.; et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In: **International Conference on Learning Representations (ICRL)**, 2020.

CLAVIJO, W., ALMEIDA, E., LOSEKANN, L., RODRIGUES, N. Impacts of the review of the Brazilian local content policy on the attractiveness of oil and gas projects, **The Journal of World Energy Law & Business** 12 (5) (2019) 449-463. doi:10.1093/jwelb/jwz030.

COLLOBERT, R.; WESTON, J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. **Proceedings of the 25th International Conference on Machine Learning**, ICML New York, USA. ACM, 2008.

CONSOLI, B., SANTOS, J., GOMES, D. *et al*. Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. in: **Proceedings of The 12th Language Resources and Evaluation Conference (LREC)**. Marseille, França, European Language Resources Association, p. 4625–4630. ISBN 979-10-95546-34-4, 2020.

CONSTANT, M., ERYIGIT, G., MONTI, J., VAN DER PLAS, L., RAMISCH, C., ROSNER, M., TODIRASCU, A. Multiword Expression Processing: A Survey, **Computational Linguistics 43** (4), 837-892. doi:10.1162/COLI_a_00302, 2017.

COLOMBO, D., PEDRONETTE, D., GUILHERME, I., PAPA, J., RIBEIRO, L., AFONSO, L., PRESOTTO, J., SOUSA, G. Discovering Patterns within the Drilling Reports using Artificial Intelligence for Operation Monitoring, **in: Offshore Technology Conference Brasil**, Rio de Janeiro, Brazil, 2019.

CORDEIRO, F. **Petrolês - Como Construir um Corpus Especializado em Óleo e Gás em Português**. PUC-Rio, Rio de Janeiro, 2020.

DELIPETREV, B., TSINARAKI, C., KOSTIĆ, U. **AI watch, historical evolution of artificial intelligence: analysis of the three main paradigm shifts in AI**. EUR 30221EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-18940-4, doi:10.2760/801580, JRC120469, 2020.

DENG, J., DONG, W., SOCHER, R., LI, L., LI, K., FEI-FEI, L. ImageNet: A large-scale hierarchical image database, **2009 IEEE Conference on Computer Vision and Pattern Recognition**, Miami, FL, pp. 248-255, doi: 10.1109/CVPR.2009.5206848. 2009

DEVLIN, J., CHANG, M.-W., LEE, K. TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, **in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171-4186, 2019.

DIAZ, F., MITRA, B., CRASWELL, N. Query Expansion with Locally-Trained Word Embeddings. in: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**, p. 367–77. doi:10.18653/v1/P16-1035, Berlin, Germany, 2016

ELMAN, J. L. Finding structure in time. **Cognitive Science**, v. 14, n. 2, p. 179–211, 1990.

ELMAN, J. L. Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. **Machine Learning - Connectionist approaches to language learning**., v. 7, n. 2–3, p. 195–225, set. 1991.

EVERINGHAM, M., ZISSERMAN, A., WILLIAMS, C. *et al.* The 2005 PASCAL Visual Object Classes Challenge. In: **Selected Proceedings of the First PASCAL Challenges Workshop, LNAI**, Springer-Verlag, 2006.

EVSUKOFF, A. **Inteligência Computacional: Fundamentos e aplicações**. Editora e-Papers, 1a edição. ISBN 978-65-8706-502-1, 2020.

FARES, M., KUTUZOV, A., OEPEN, S., VELLDAL, E. Word vectors, reuse, and replicability: Towards a community repository of large-text resources, **in: Proceedings of the 21st Nordic Conference on Computational Linguistics**, Association for Computational Linguistics, Gothenburg, Sweden, pp. 271-276, 2017.

FARUQUI, M., DODGE, J., JAUHAR, S., DYER, C., HOVY, E., SMITH, N. Retrofitting word vectors to semantic lexicons, in: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Association for Computational Linguistics, Denver, Colorado, pp. 1606-1615, doi:10.3115/v1/N15-1184. 2015.

FARUQUI, M., TSVETKOV, Y., RASTOGI, P., DYER, C. Problems with evaluation of word embeddings using word similarity tasks, **in: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP**, Association for Computational Linguistics, Berlin, Germany, pp. 30-35, doi:10.18653/v1/W16-2506. 2016.

FERNÁNDEZ, E., PEDROSA, O., PINHO, A. **Dicionário do petróleo em língua portuguesa**, Rio de Janeiro, RJ. Lexikon : PUC Rio, ISBN: 978-8586368554. 2009.

FORBES, 2017. **The Big (Unstructured) Data Problem**. Online. Disponível em> <https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#67629e6e493a>. Acesso em 10/12/2020.

FURTADO, P. **Interpretação automática de relatórios de operação de equipamentos**. Dissertação de Mestrado. Pontifícia Universidade Católica Do Rio De Janeiro. doi:10.17771/PUCRio.acad.30732, 2017.

GARTNER, 2011. **Information Management Goes 'Extreme': The Biggest Challenges for 21st Century CIOs.** [online]. Disponível em: <http://togetherwepass.co.za/http://togetherwepass.co.za/wp-content/uploads/filebase/certified_office_manager/management/study_notes_/Extreme-Information-Management%20Extreme.pdf>. Acesso em março/2019

GERS, F., SCHMIDHUBER, J., CUMMINS, F. Learning to forget: Continual prediction with LSTM, **9th International Conference on Artificial Neural Networks**, pp. 850–855, 1999.

GLADKOVA, A., DROZD, A. Intrinsic Evaluations of Word Embeddings: What Can We Do Better?, **in: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP**, Association for Computational Linguistics, Berlin, Germany, pp. 36-42, 2016.

GOLDBERG, Y. (2016). A primer on neural network models for natural language processing. **Journal of Artificial Intelligence Research**, 57, 345-420.

GOMES, D., CORDEIRO, F., EVSUKOFF, A. 2018. Word Embeddings em Português para o Domínio Específico de Óleo e Gás. **In: Proceedings of Rio Oil & Gas Expo and Conference 2018**.

GOMES, D., CORDEIRO, F.;, CONSOLI, B., SANTOS, N., MOREIRA, V., VIEIRA, R., MORAES, S., EVSUKOFF, A. Portuguese word embeddings for the oil and gas industry: Development and evaluation. **Computers in Industry**, v. 124, p. 103347. https://doi.org/10.1016/j.compind.2020.103347. 2021

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. **Deep Learning**. MIT Press. Disponivel em: <http://www.deeplearningbook.org>, 2016.

HARTMAN, N. S., FONSECA, E., SHULBY, C., TREVISO, M., RODRIGUES, J. S., ALUISIO, S. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. in: **Proceedings of Symposium in Information and Human Language Technology**. Uberlandia, MG, Sociedade Brasileira de Computacão, 2017.

HARRIS, Z. Distributional Structure, **WORD 10** (2-3), 146-162, 1954.

HE, K., ZHANG, X., REN S., SUN, J. Deep Residual Learning for Image Recognition, **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90, 2016.

HENRIETTE, E.; FEKI, M.; BOUGHZALA, I. The Shape of Digital Transformation: A Systematic Literature Review. **MCIS 2015 Proceedings**, 2015. Disponível em: <https://aisel.aisnet.org/mcis2015/10>. Acesso e 01/11/2020

HEWITT, J., MANNING, C. A structural probe for finding syntax in word representations. **in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Volume 1, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419, 2019.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997.

HOWARD, J., RUDER, S. Universal Language Model Fine-tuning for Text Classification. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, p. 328–339, 2018.

ITTOO, A.; NGUYEN, L. M.; VAN DEN BOSCH, A. Text analytics in industry: Challenges, desiderata and trends. **Computers in Industry**, Natural Language Processing and Text Analytics in Industry. v. 78, p. 96–107, 1 maio 2016.

JACOBS, T. The Oil and Gas Chat Bots Are Coming, **Journal of Petroleum Technology 71** (02), 34-36, 2019.

JIANG, Z., LI, L., HUANG, D. , JIN, L. Training word embeddings for deep learning in biomedical text mining tasks, **in: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**, pp. 625-628, 2015.

JOLLIFFE, I. **Principal Component Analysis and Factor Analysis**, pages 115–128. Springer New York, NY, 1986.

JOZEFOWICZ, R., ZAREMBA, W., SUTSKEVER, I. An Empirical Exploration of Recurrent Network Architectures. In: **Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37**. [s.l.]: JMLR.org, 2015, p. 2342–2350. (ICML'15).

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. Prentice Hall Series in Artificial Intelligence. 3. ed. Prentice Hall, 2020.

KALYAN, K., SANGEETHA, S. SECNLP: A survey of embeddings in clinical natural language processing, **Journal of Biomedical Informatics** 101, 103323, 2020.

KHABIRI, E., GIFFORD, W., VINZAMURI, B., PATEL, D., MAZZOLENI, P. Industry Specific Word Embedding and its Application in Log Classification, **in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management**, CIKM '19, Association for Computing Machinery, Beijing, China, pp. 2713i2721, 2019.

KHURANA, Diksha; KOLI, Aditya; KHATTER, Kiran., SINGH, S. (2017). **Natural Language Processing: State of The Art, Current Trends and Challenges**. *CoRR abs/1708.05148*.

KIM, Y. Convolutional Neural Networks for Sentence Classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (Emnlp),** 2014

KITAEV, N., KAISER, L., LEVSKAYA, A. Reformer: The Efficient Transformer. In: **International Conference on Learning Representations (ICRL)**, 2020.

KOWSARI, K., MEIMANDI, K., HEIDARYSAFA, M., MENDU, S., BARNES, L., BROWN, D. (2019) Text Classification Algorithms: A Survey. **Information 10**, nº 4 (abril de 2019): 150. https://doi.org/10.3390/info10040150.

KRIZHEVSKY, A. **Learning multiple layers of features from tiny images**, 2009. Disponível em: < https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. Acesso em: 02/2021.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **ImageNet Classification with Deep Convolutional Neural Networks**. Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. NIPS'12.USA, 2012.

LAI, S., LIU, K., XU, L., ZHAO, J. How to Generate a Good Word Embedding. **IEEE Intelligent Systems**, vol. 31, no. 6, pp. 5-14. doi: 10.1109/MIS.2016.45. Nov.-Dec. 2016.

LAMPLE, G.; BALLESTEROS, M.; SUBRAMANIAN, S., KAWAKAMI, K., DYER, C**.** Neural Architectures for Named Entity Recognition. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: Association for Computational Linguistics, 2016, p. 260–270. Disponível em: <http://aclweb.org/anthology/N16-1030>. Acesso em: 9 abr. 2019.

LE, Q., MIKOLOV, T. Distributed Representations of Sentences and Documents. In: **International Conference on Machine Learning**. [s.l.: s.n.], 2014.

LECUN, Y., BENGIO, Y., HINTON, G. Deep Learning. **Nature**, 2015.

LEE, J., YOON, W., KIM, S., KIM, D., KIM, S., SO, C., KANG, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining, **Bioinformatics** 36 (4) (2020) 1234-1240.

LEPIKHIN, D., LEE, H., XU, Y., *et al*. GShard: **Scaling Giant Models with Conditional Computation and Automatic Sharding**. arXiv:2006.16668, 2020.

LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. **Transactions of the Association for Computational Linguistics**, v. 3, p. 211–225, 2015.

LIKERT, R. A technique for the measurement of attitudes. **Archives of Psychology**, 22 140, p. 55-55, 1932.

LINZEN, T. Issues in evaluating semantic spaces using word analogies, **in: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP**, Association for Computational Linguistics, Berlin, Germany, pp. 13-18. doi:10.18653/v1/W16-2503, 2016.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., STOYANOV, V. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. ArXiv abs/1907.11692, 2019

LOPER, E., BIRD, S. NLTK: the Natural Language Toolkit. **Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1** (ETMTNLP '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117, 2002

LU, H., GUO, L., AZIMI, M., HUANG, K. Oil and Gas 4.0 era: A systematic review and outlook, **Computers in Industry** 111, p. 68-90, 2019.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge, Mass: MIT Press, 1999.

MANNING. **Natural language processing with deep learning cs224n: Lecture 14: More on Contextual Word Representations and Pretraining**. Notas de Aula, 2020. Disponível em: <https://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture14-contextual-representations.pdf>. Acesso em 11/01/2020

MANNING, C. Computational Linguistics and Deep Learning. **Computational Linguistics**, MIT Press v. 41, n. 4, p. 701–707, doi: 10.1162/COLI_a_00239, 2015.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. New York: Cambridge University Press, 2008.

MANNING, C., SOCHER, R. **Lecture Notes on Natural Language Processing with Deep Learning**. Stanford University, 2017. Disponível em: <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/syllabus.html>

MARQUES, J., COZMAN, F., SANTOS, I. Automatic Summarization of Technical Documents in the Oil and Gas Industry, **in: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**, pp. 431-436, iSSN: 2643-6264, 2019.

MATT, Christian; HESS, Thomas; BENLIAN, Alexander. Digital Transformation Strategies. **Business & Information Systems Engineering**, v. 57, n. 5, p. 339–343, 2015.

MATTMANN, C., ZITTING, J. **Tika in action**. Manning Publications Co., 2011

MICROSOFT. **Turing-NLG: A 17-billion-parameter language model by Microsoft**, 2020. Disponível em: <https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>. Acesso em: 12 jan. 2021.

MIKOLOV, T., CHEN, K., CORRADO, G., DEAN, J. Efficient estimation of word representations in vector space. **ICLR Workshop**, 2013a.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G., DEAN, J. Distributed representations of words and phrases and their compositionality. **Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2(NIPS'13)**. 2013b.

MIKOLOV, T., KARAFIÁT, M., BURGET, L., CERNOCKÝ, J., KHUDANPUR, S. (2010). Recurrent neural network based language model. **Proceedings of the 11th Annual Conference of the International Speech Communication Association**, INTERSPEECH. 2. 1045-1048, 2010.

MISHRA, S., SHARMA, A. On the Use of Word Embeddings for Identifying Domain Specific Ambiguities in Requirements, **in: 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)**, pp. 234-240, 2019.

MOOSAVI, N., FAN, A., SHWARTZ, V., *et al* (Orgs.). **Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing**. Online: Association for Computational Linguistics, 2020. Disponível em: <https://www.aclweb.org/anthology/2020.sustainlp-1.0>. Acesso em: 12 jan. 2021

MUNEEB, T. H., SAHU, S. K., ANAND, A. Evaluating distributed word representations for capturing semantics of biomedical concepts. **Workshop on Biomedical Natural Language Processing (BioNLP).** 2015.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association: JAMIA**, v. 18, n. 5, p. 544–551, out. 2011.

NEELAKANTAN, A. SHANKAR, J.; PASSOS, A. McCallum, A. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).** Doha, Qatar: Association for Computational Linguistics, 2014

NEWMAN-GRIFFIS, D. LAI, A. FOSLER-LUSSIER, E. Insights into analogy completion from the biomedical domain, **in: Workshop on Biomedical Natural Language Processing (BioNLP)**, Association for Computational Linguistics, Vancouver, Canada, pp. 19-28. doi:10.18653/v1/W17-2303, 2017

NEWMAN-GRIFFIS, D. LAI, A. FOSLER-LUSSIER, E Jointly embedding entities and text with distant supervision, **in: Proceedings of The Third Workshop on Representation Learning for NLP**, Association for Computational Linguistics, Melbourne, Australia, pp. 195-206. doi: 10.18653/v1/W18-3026, 2018.

NIKLAUS, C., CETTO, M., FREITAS, A., HANDSCHUH, S. **A survey on open information extraction**. CoRR, abs/1806.05599, 2018.

NOORALAHZADEH, F., ØVRELID, L., LØNNING, J. 2018. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

NOORALAHZADEH, F. **Low-Resource Adaptation of Neural NLP Models**, P.hD. thesis, Faculty of Mathematics and Natural Sciences, University of Oslo, 2020.

OLAH, C. **Understanding LSTM Networks**. 2015 Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Acesso em: abr, 2019.

PADARIAN, J., FUENTES, I. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts, **SOIL**, 177-187, 2019.

PAKHOMOV, S., FINLEY, G., MCEWAN, R., WANG, Y, MELTON, G. Corpus domain effects on distributional semantic modeling of medical terms, **Bioinformatics**, 2016.

PENNINGTON, J, SOCHER, R., MANNING, C. D. Glove: Global vectors for word representation, **in Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1532–1543.http://www.aclweb.org/anthology/D14-1162, 2014.

PERRAULT, R., SHOHAM, Y., BRYNJOLFSSON, E., CLARK, J., ETCHEMENDY, J., GROSZ, B., LYONS, T., MANYIKA, J., MISHRA, S., NIEBLES, J. **The AI Index 2019 Annual Report**, AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, dezembro 2019.

PETERS, M., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K., ZETTLEMOYER, L. 2018. Deep contextualized word representations. **Proceedings of NAACL 2018**.

PILEHVAR, M., COLLIER, N. Improved semantic representation for domain-specific entities, **in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing**, Association for Computational Linguistics, Berlin, Germany, pp. 12-16. doi:10.18653/v1/W16-2902, 2016.

PILEHVAR, M., CAMACHO-COLLADOS, J. **Embeddings in Natural Language Processing**. Morgan & Claypool Publishers, San Rafael, CA, 2020.

POLIGNANO, M., GEMMIS, M., SEMERARO, G. Contextualized BERT sentence embeddings for author profiling: The cost of performances, **in: Computational Science and Its Applications - ICCSA 2020**, Springer International Publishing, Cham, pp. 135-149, 2020.

RADFORD, A., NARASIMHAN, K., SALIMANS, T., SUTSKEVER, I. **Improving language understanding with unsupervised learning**. In: Technical report, OpenAI, 2018. Disponível em: <https://openai.com/blog/language-unsupervised/>. Acesso em 05/01/2021.

RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I. **Improving language understanding with unsupervised learning**. In: Technical report, OpenAI, 2019. Disponível em: <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf/>. Acesso em 001/2021.

RAJPURKAR, P. ZHANG, J.; LOPYREV, K., LIANG, P**.** SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**.. https://doi.org/10.18653/v1/D16-1264, 2016.

RÉ, C., NIU, F., GUDIPATI, P., SRISUWANANUKORN, C. Overton: A data system for monitoring and improving machine-learned products. In: **10th Annual Conference on Innovative Data Systems Research (CIDR 20)**, Amsterdam, Netherlands, 2020.

REHUREK, R., SOJKA, P. **Gensim–python framework for vector space modelling**. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 2011.

RIBEIRO, L., AFONSO, L., COLOMBO, D., GUILHERME, I., PAPA, J. Evolving Neural Conditional Random Fields for drilling report classification, **Journal of Petroleum Science and Engineering 187,** 2020.

RODRIGUES, J., BRANCO, A., NEALE, S., SILVA, J. LX-DSemVectors: Distributional Semantics Models for Portuguese. **Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016).** Springer International Publishing, 2016.

RODRIGUES, J., BRANCO, A. Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese. **In: Proceedings of the 11th International Conference on Language Resources and Evaluation**, 2018.

ROGERS, A., DROZD, A., LI, B. The (too many) problems of analogical reasoning with word vectors, **in: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)**, Association for Computational Linguistics, Vancouver, Canada, pp. 135-148. doi:10.18653/v1/S17-1017, 2017

ROGERS, A; DROZD, A; RUMSHISKY, A; et al (Orgs.). **Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP**. Minneapolis, USA: Association for Computational Linguistics, 2019. Disponível em: <https://www.aclweb.org/anthology/W19-2000>. Acesso em: 15 jan. 2021.

ROGERS, A., KOVALEVA, O. RUMSHISKY, A. A Primer in BERTology: What We Know About How BERT Works. **Transactions of the Association for Computational Linguistics,** v. 8, p. 842–866, doi:10.1162/tacl_a_00349, 2020.

RUDER, S. **NLP's ImageNet moment has arrived.** 2018**.** Disponível em: <http://ruder.io/nlp-imagenet/>. Acesso em: 9 abr. 2019.

RUDER, S. **Neural Transfer Learning for Natural Language Processing**. Ph.D. thesis. National University of Ireland, Galway, 2019.

RUDER, S., VULIC, I., SØGAARD, A. A Survey of Cross-Lingual Word Embedding Models. **The Journal of Artificial Intelligence Research**. doi: 10.17863/CAM.30462, 2019.

RUMELHART, D., HINTON, G., WILLIAMS, R. Learning representations by back-propagating errors, **Nature** 323 (6088) 533-536, 1986.

SAHLGREN, M. The Distributional Hypothesis. **Rivista di Linguistica** (Italian Journal of Linguistics), 20. pp. 33-53, 2008

SANCHEZ-PI N., MARTÍ L., GARCIA B. Text Classification Techniques in Oil Industry Applications. In: International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. **Advances in Intelligent Systems and Computing**, vol 239. Springer, Cham. doi: 10.1007/978-3-319-01854-6_22, 2014.

SANH, V., DEBUT, L., CHAUMOND, J., WOLF, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. in: **5th Workshop on Energy Efficient Machine Learning and Cognitive Computing – NeurIPS**, 2019

SANTOS, J. CONSOLI, B., SANTOS, C., TERRA, J., COLLOVINI, S., VIEIRA, R. Assessing the impact of contextual embeddings for portuguese named entity recognition, **in: Proceedings of the 8th Brazilian Conference on Intelligent Systems**, IEEE, pp. 437-442. doi:10.1109/BRACIS.2019.00083, 2019.

SCHNABEL, T., LABUTOV, I., MIMNO, D., JOACHIMS, T. Evaluation methods for unsupervised word embeddings. in: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, Association for Computational Linguistics, Lisbon, Portugal, pp. 298-307, 2015.

SCHWARTZ, R., DODGE, J., SMITH, N., *et al.* Green AI. **Communications of the ACM**, v. 63, n. 12, p. 54–63, 2020.

SHALABY, W., ZADROZNY, W. Mined semantic analysis: A new concept space model for semantic representation of textual data, **in: 2017 IEEE International Conference on Big Data (Big Data)**, pp. 2122-2131, 2017.

SHOHAM, Y., PERRAULT, R., BRYNJOLFSSON, E., CLARK, J., MANYIKA, J., NIEBLES, J., LYONS, T., ETCHEMENDY, J., GROSZ, B., BAUER, Z, **The AI Index 2018 Annual Report**, AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA, December 2018.

SIMONYAN, K., ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. arXiv:1409.1556 [cs], 2015

SMILKOV, D., THORAT, N., NICHOLSON, C., *et al*. **Embedding Projector: Interactive Visualization and Interpretation of Embeddings**, arXiv:1611.05469, 2016.

SOCHER, R.; PERELYGIN, A.; WU, J., CHUANG, J., MANNING, C., NG, A., POTTS, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, p. 1631–1642, 2013.

STEIGER, J. Tests for comparing elements of a correlation matrix, **Psychological bulletin** 87 (2), 1980.

STRUBELL, E., GANESH, A., MCCALLUM, A. Energy and Policy Considerations for Deep Learning in NLP. **in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, p. 3645–3650. doi: 10.18653/v1/P19-1355, 2019.

SUTSKEVER, Ilya; VINYALS, Oriol; LE, Quoc V. Sequence to Sequence Learning with Neural Networks. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2014, p. 3104–3112. (NIPS'14).

TSHITOYAN, V., DAGDELEN, J., WESTON, L., DUNN, A., RONG, Z., KONONOVA, O., PERSSON, K., CEDER, G., JAIN, A. Unsupervised word

embeddings capture latent knowledge from materials science literature, **Nature 571** (7763) 95-98, 2019.

TURIAN, J., RATINOV, L., BENGIO, Y. Word Representations: A Simple and General Method for Semi-Supervised Learning, **in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**, Association for Computational Linguistics, Uppsala, Sweden, pp. 384-394, 2010.

TURNEY, P., PANTEL, P. From Frequency to Meaning: Vector Space Models of Semantics, **Journal of Artificial Intelligence Research,** 37, 141-188, 2010.

VAN DER MAATEN, L.J.P., HINTON, G. Visualizing High-Dimensional Data Using t-SNE. **Journal of Machine Learning Research 9**: pp. 2579-2605, 2008.

VASWANI, A. SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A., KAISER, L., POLOSUKHIN, I. Attention is All you Need. In: **Advances in Neural Information Processing Systems 30**. [s.l.] Curran Associates, Inc. p. 5998–6008, 2017.

WANG, Y., LIU, S., AFZAL, N., RASTEGAR-MOJARAD, M., WANG, L., SHEN, F., KINGSBURY, P., LIU, H. A comparison of word embeddings for the biomedical natural language processing, **Journal of Biomedical Informatics 87**, 12-20, 2018.

WANG, C., MA, X., CHEN, J., CHEN, J. Information extraction and knowledge graph construction from geoscience literature, **Computers & Geosciences 112**, 112-120, 2018a.

WU, Y., SCHUSTER, M., CHEN, Z., *et al.* **Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.** CoRR, abs/1609.08144.

YADAV, V., BETHARD, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. **Proceedings of the 27th International Conference on Computational Linguistics**, pages 2145–2158 Santa Fe, New Mexico, USA, August 20-26, 2018.

YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R., LE, Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. **In: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019**, **NeurIPS 2019**. Curran Associates, Inc., pp. 5754–5764, 2019.

YIN, W.; KANN, K.; YU, M.; SCHÜTZE, H. **Comparative Study of CNN and RNN for Natural Language Processing.** arXiv:1702.01923, 2017.

YOUNG, T., HAZARIKA, D., PORIA, S., CAMBRIA, E. Recent Trends in Deep Learning Based Natural Language Processing. **IEEE Computational Intelligence Magazine**, vol.13. doi: 10.1109/MCI.2018.2840738, 2018

UCHEREK, J., LAWAL, T., PRINZ, M., LI, L., ASHOK, P., VAN OORT, E., GOBERT, T., MEJIA, J. **Auto-Suggestive Real-Time Classification of Driller Memos into Activity Codes Using Natural Language Processing**, Society of Petroleum Engineers, 2020.

NILC. **Repositório de Word Embeddings do NILC**. NILC - Núcleo Interinstitucional de Linguística Computacional. Disponível em: <http://www.nilc.icmc.usp.br/embeddings>. Acesso em 05/05/2020.

WANG, A., SINGH, A., MICHAEL, J., HILL, F., LEVY, O., BOWMAN, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. **in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.** Brussels, Belgium: Association for Computational Linguistics, p. 353–355, 2018.

WANG, B., WANG, A., CHEN, F., WANG, Y., KUO, C. Evaluating Word Embedding Models: Methods and Experimental Results. **APSIPA Transactions on Signal and Information Processing 8**. doi:10.1017/ATSIP.2019.12, 2019.

WANG, S., LI, B., KHABSA, M., *et al*. **Linformer: Self-Attention with Linear Complexity.** arXiv:2006.04768, 2020.

WILSON, A. Natural-Language-Processing Techniques for Oil and Gas Drilling Data, **Journal of Petroleum Technology 69** (10) 96-97, 2017.

WOLF, T., *et al*, Transformers: State-of-the-Art Natural Language Processing, in: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, Online: Association for Computational Linguistics, p. 38–45, 2020.

WORLD ECONOMIC FORUM. **Digital Transformation Initiative: Oil and Gas Industry.** 2017 [White paper]. Disponível em: < http://reports.weforum.org/digital-transformation/oil-gas//>. Acesso em: 15 abr. 2019.

XIA, P., WU, S., VAN DURME, B. Which *BERT? A Survey Organizing Contextualized Encoders. **in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, p. 7516–7533. doi: 10.18653/v1/2020.emnlp-main.608, 2020.

ZEILER M., FERGUS R. Visualizing and Understanding Convolutional Networks. In: **Lecture Notes in Computer Science**, vol 8689. Springer, https://doi.org/10.1007/978-3-319-10590-1_53, 2014.

ZHANG, Z., GENTILE, A., CIRAVEGNA, F. Recent advances in methods of lexical semantic relatedness - a survey, **Natural Language Engineering** 19 (4), 411-479, 2013.

ZHANG, L., WANG, S., LIU, B. (2018). **Deep Learning for Sentiment Analysis: A Survey**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 10.1002/widm.1253, 2018

ZHENZHONG, L., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P., SORICUT, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. **International Conference on Learning Representations**, 2020.

Apêndice

A.1. Categorias de entidades no GeoCorpus

Tabela 5.1: Comparação do detalhamento das categorias, considerando a versão Original e a versão revisada do GeoCorpus
Fonte: VIEIRA *et al.* (2020)[59]

Classe	Instâncias (Original)	Instâncias (Revisada)
Tempo		
idade	796	799
eon	288	256
era	326	414
epoca	650	687
periodo	637	714
Rochas		
metamorficas	197	378
magmaticas	222	582
sedimentaresSiliciclasticas	741	1102
sedimentaresCarbonaticas	240	355
sedimentaresQuimicas	5	12
sedimentaresOrganicas	22	22
Constituintes e Propriedades de Rochas		
constituinteRochaSedimentar	0	112
mineral	0	212
fosseis	0	132
estruturaSedimentar	0	86
estruturaGeologica	0	78
Sítio		
contextoGeologicoDeBacia	262	663
ambienteSedimentacao	0	146
bentonico	13	27
planctonico	44	112
campoPetrolifero	0	6
Elementos da Estratigrafia		
baciaSedimentar	243	552
unidadeEstratigrafica	578	764

[59] VIEIRA *et al., 2020.* GeoCorpus Change Report. Disponível no repositório público PetroVec: https://github.com/Petroles/Petrovec/blob/master/GeoCorpus%20V3/GeoCorpus%20Change%20Report.pdf

unidadeGeotectonica	0	28
estratigrafia	0	247
formacao	18	0
Outros		
sistemaPetrolifero	0	93
estruturaDeBacia	40	0
geomorfologia	0	54
granulometria	67	129
elementoQuimico	0	26
procedimentoMetodologico	0	166
outro	737	0
Soma	***6.126***	***8.954***

A.2. Planos de Desenvolvimento dos Campos - ANP

Tabela 5.2: Listagem completa dos campos. Fonte: ANP[60]

Campos de Petróleo
Abalone, Acajá-Burizinho, Acauã, Água Grande, Aguilhada, Agulha, Albacora, Albacora Leste, Albatroz, Alto do Rodrigues, Anambé, Andorinha, Anequim, Angelim, Angico, Apraiús, Arabaiana, Araçari, Araçás, Araçás Leste, Aracuã, Arapaçu, Arara Azul, Araracanga, Aratu, Aratum, Argonauta, Arribaçã, Aruari, Asa Branca, Atalaia Sul, Atlanta, Atum, Azulão
Badejo, Bagre, Baixa do Algodão, Baixa do Juazeiro, Barracuda, Barrinha, Barrinha Leste, Barrinha Sudoeste, Baúna, Barra Bonita, Barra do Ipiranga, Bem-te-vi, Benfica, Bicudo, Biguá, Bijupirá, Biquara, Biriba, Boa Esperança, Boa Vista, Bom Lugar, Bonito, Bonsucesso, Brejinho, Brejo Grande, Buracica, Búzios
Cachoeirinha, Cacimbas, Caioba, Camarupim, Camarupim Norte, Cambacica, Camorim, Campo Grande, Canabrava, Canapu, Canário, Cancã, Candeias, Cangoá, Cantagalo, Canto do Amaro, Carapanaúba, Carapeba, Carapiá, Carapicu, Carapitanga, Carataí, Caratinga, Carcará, Cardeal, Cardeal Amarelo, Cardeal do Nordeste, Carmópolis, Cassarongongo, Castanhal, Catuá, Cexis, Chauá, Cherne, Cidade de Aracaju, Cidade de Entre Rios, Cidade de São Miguel dos Campos, Cidade de Sebastião Ferreira, Cioba, Colibri, Conceição, Concriz, Congro, Coqueiro Seco, Coral, Córrego Cedro Norte, Córrego Cedro Norte Sul, Córrego das Pedras, Córrego Dourado, Corvina, Crejoá, Cupiúba, Curimã
Dom João, Dom João Mar, Dó-ré-mi, Dourado
Enchova, Enchova Oeste, Espada, Espadarte, Estreito
Fazenda Alegre, Fazenda Alto das Pedras, Fazenda Alvorada, Fazenda Azevedo, Fazenda Bálsamo, Fazenda Belém (BA), Fazenda Belém (RN), Fazenda Boa Esperança, Fazenda Canaan, Fazenda Cedro, Fazenda Cedro Norte, Fazenda Curral, Fazenda Guindaste, Fazenda Imbé, Fazenda Junco, Fazenda Malaquias, Fazenda Matinha, Fazenda Onça, Fazenda Panelas, Fazenda Pau Brasil, Fazenda Pocinho, Fazenda Queimadas, Fazenda Rio Branco, Fazenda Santa Luzia, Fazenda Santa Rosa, Fazenda Santo Estevão, Fazenda São Jorge, Fazenda São Rafael, Foz do Vaza-Barris, Frade, Furado
Gaivota, Galo de Campina, Garoupa, Garoupinha, Gavião Azul, Gavião Branco, Gavião Branco Norte, Gavião Preto, Gavião Real, Gavião Vermelho, Golfinho, Gomo, Graúna, Guaiuba, Guamaré, Guamaré Sudeste, Guanambi, Guará, Guriatã, Guriri
Harpia

[60] http://www.anp.gov.br/exploracao-e-producao-de-oleo-e-gas/gestao-de-contratos-de-e-p/fase-de-producao/planos-de-desenvolvimento. Acesso em 01/07/2020

Icapuí, Ilha de Bimbarra, Ilha Pequena, Inhambu, Iraúna, Irerê, Itaparica, Itapu
Jacuípe, Jacupemba, Jacutinga, Jandaia, Jandaia Sul, Jaçanã, Janduí, Japiim, Japuaçu, Jequiá, Jiribatuba, João de Barro, Juazeiro, Jubarte, Juriti, Juruá
Lagoa Aroeira, Lagoa Bonita, Lagoa Pacas, Lagoa Parda, Lagoa Parda Norte, Lagoa do Paulo, Lagoa do Paulo Norte, Lagoa do Paulo Sul, Lagoa Piabanha, Lagoa Suruaca, Lagosta, Lamarão, Lapa, Leodório, Leste de Poço Xavier, Leste do Urucu, Linguado, Livramento, Lorena, Lula
Macau, Maçarico, Malhado, Malombê, Manati, Mandacaru, Mapele, Marimbá, Mariricu, Mariricu Norte, Mariricu Oeste, Maritaca, Marlim, Marlim Leste, Marlim Sul, Massapê, Massuí, Mata de São João, Mato Grosso, Mãe-da-lua, Merluza, Mexilhão, Miranga, Miranga Norte, Monte Alegre, Moréia, Morrinho, Morro do Barro, Mossoró, Mutum
Namorado, Nativo Oeste, Nordeste de Namorado, Norte de Fazenda Caruaçu
Oeste de Ubarana, Oliva, Ostra
Pajeú, Papa-terra, Parati, Pardal, Pargo, Pariri, Pajeú, Pampo, Paru, Patativa, Paturi, Pedra Sentada, Pedrinhas, Peregrino, Periquito, Periquito Norte, Peroá, Pescada, Pilar, Pinaúna, Piranema, Pintassilgo, Piracucá, Piraúna, Pitangola, Pitiguari, Pojuca, Poço Verde, Poço Xavier, Polvo, Ponta do Mel, Porto Carão
Quererá
Rabo Branco, Redonda, Redonda Profundo, Remanso, Riacho da Barra, Riacho da Forquilha, Riacho Ouricuri, Riacho São Pedro, Riacho Velho, Riachuelo, Rio Barra Seca, Rio da Serra, Rio do Bu, Rio dos Ovos, Rio Ipiranga, Rio Itariri, Rio Itaúnas, Rio Itaúnas Leste, Rio Joanes, Rio Mossoró, Rio Pipiri, Rio Pojuca, Rio Preto, Rio Preto Oeste, Rio Preto Sul, Rio São Mateus, Rio São Mateus Oeste, Rio Sauípe, Rio Subaúma, Rio Urucu, Rolinha, Roncador
Sabiá, Sabiá Bico-de-Osso, Sabiá da Mata, Saíra, Salema, Salgo, Salina Cristal, Sanhaçu, Santana, São Manoel, São Mateus, São Mateus Leste, São Miguel dos Campos, São Pedro, Sapinhoá, Sebastião Ferreira, Sempre Viva, Sépia, Seriema, Serra, Serra do Mel, Serra Vermelha, Serraria, Sesmaria, Sibite, Siri, Siririzinho, Socorro, Socorro Extensão, Sudoeste Urucu, Sul de Coruripe, Sul de Lula, Sussuarana
Tabuiaiá, Tabuleiro dos Martins , Tambaú, Tambuatá, Tangará, Tapiranga, Tapiranga Norte, Taquipe, Tartaruga, Tartaruga Verde, Tatuí, Tico-tico, Tiê, Tigre, Tiziu, Três Marias, Trilha, Trinca Ferro, Trovoada, Tubarão Azul, Tubarão Martelo, Tucano
Ubarana, Uirapuru, Uirapuru Sudoeste, Upanema, Uruguá, Urutau
Varginha, Vermelho, Viola, Voador
Xaréu