Universidade de Brasília

Faculdade de Tecnologia

Departamento de Engenharia Elétrica

Similaridade Semântica Textual em Língua Portuguesa: Estudo Comparativo de Modelos Multilíngues e Monolíngues

Roberto Caixeta

TRABALHO DE CONCLUSÃO DE CURSO

ENGENHARIA DE COMPUTAÇÃO

Brasília

2025

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Elétrica

Similaridade Semântica Textual em Língua Portuguesa: Estudo Comparativo de Modelos Multilíngues e Monolíngues

Roberto Caixeta

Trabalho de Conclusão de Curso submetido como requisito parcial para obtenção do grau de Engenheiro de Computação.

Orientador: Prof. Dr. Daniel Guerreiro e Silva

Brasília
2025

FICHA CATALOGRÁFICA

Caixeta, Roberto.

Similaridade Semântica Textual em Língua Portuguesa: Estudo Comparativo de Modelos Multilíngues e Monolíngues / Roberto Caixeta; orientador Daniel Guerreiro e Silva. -- Brasília, 2025.

50 p.

Trabalho de Conclusão de Curso (Engenharia de Computação) -- Universidade de Brasília, 2025.

1. Similaridade Semântica Textual. 2. Processamento de Linguagem Natural. 3. S-BERT. 4. Fine-Tuning. I. Silva, Daniel Guerreiro e, orient. II. Título.

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Elétrica

Similaridade Semântica Textual em Língua Portuguesa: Estudo Comparativo de Modelos Multilíngues e Monolíngues

Roberto Caixeta

Trabalho de Conclusão de Curso submetido como requisito parcial para obtenção do grau de Engenheiro de Computação.

Trabalho aprovado. Brasília, 12 de fevereiro de 2025:

Prof. Dr. Daniel Guerreiro e Silva,
UnB/FT/ENE
Orientador

Profa. Dra. Cláudia Jacy Barenco Abbas,
UnB/FT/ENE
Examinador interno

Prof. Dr. Alexandre Ricardo Soares Romariz,
UnB/FT/ENE
Examinador interno

Agradecimentos

Em primeiro lugar, gostaria de agradecer e dedicar este trabalho ao meu pai, Roberto, e às minhas irmãs, Marina e Nayara, por terem me proporcionado todo o apoio e suporte durante a minha graduação. Também dedico este trabalho a todos os meus amigos de graduação que me acompanharam durante esse período. Além disso, gostaria de agradecer a todos os professores que tive ao longo da minha vida, pois os ensinamentos de cada um contribuíram para a realização deste trabalho, em especial ao meu orientador, Daniel Guerreiro e Silva, por todo o comprometimento e paciência durante o processo.

"If you find that you're spending almost all your time on theory,
start turning some attention to practical things;
it will improve your theories.
If you find that you're spending almost all your time on practice,
start turning some attention to theoretical things;
it will improve your practice."
(Donald Knuth)

Resumo

Este trabalho apresenta uma análise da Similaridade Semântica Textual (STS) na língua portuguesa, avaliando o desempenho de modelos multilinguísticos e monolíngues em corpora específicos, como ASSIN, ASSIN 2 e SICK-BR. São exploradas metodologias de pré-treinamento e *fine-tuning*, com o objetivo de adaptar modelos como XLM-RoBERTa, LaBSE, MiniLM-L12-v2 e BERT-base-portuguese às particularidades linguísticas do português. Os experimentos realizados demonstraram que a aplicação do *fine-tuning* melhora o desempenho dos modelos, resultando em escores mais alinhados às anotações humanas nas métricas utilizadas.

Os resultados obtidos indicam que, embora o BERT-base-portuguese, por ser um modelo monolíngue especializado no idioma, apresente bom desempenho, sua performance foi ligeiramente inferior a modelos multilinguísticos, como o LaBSE e o MiniLM-L12-v2, que se beneficiam do treinamento em um volume de dados mais amplo e diversificado. Esse estudo também ressalta a importância dos corpora em português existentes, evidenciando seu papel na adaptação e avaliação dos modelos para a tarefa de STS.

A aplicação prática da Similaridade Semântica Textual é destacada em diversos contextos, como sistemas jurídicos, educação e saúde, demonstrando seu potencial para aprimorar aplicações que exigem interpretação precisa de textos em língua portuguesa. Por fim, este estudo contribui para o avanço acadêmico e tecnológico ao propor soluções para o processamento de textos em língua portuguesa.

Palavras-chave: Similaridade Semântica Textual; Processamento de Linguagem Natural; S-BERT; Fine-Tuning.

Abstract

This study presents an analysis of Semantic Textual Similarity (STS) in the Portuguese language, evaluating the performance of both multilingual and monolingual models on specific corpora, such as ASSIN, ASSIN 2, and SICK-BR. The study explores pre-training and *fine-tuning* methodologies, aiming to adapt models such as XLM-RoBERTa, LaBSE, MiniLM-L12-v2, and BERT-base-portuguese to the linguistic particularities of Portuguese. The experiments conducted demonstrated that applying *fine-tuning* improves model performance, resulting in scores more aligned with human annotations according to the evaluation metrics used.

The results indicate that, although BERT-base-portuguese, being a monolingual model specifically designed for the language, performs well, its performance was slightly inferior to multilingual models such as LaBSE and MiniLM-L12-v2, which benefit from training on a broader and more diverse dataset. This study also highlights the importance of existing Portuguese corpora, emphasizing their role in adapting and evaluating models for the STS task.

The practical application of Semantic Textual Similarity is emphasized in various fields, such as legal systems, education, and healthcare, demonstrating its potential to enhance applications that require precise interpretation of texts in Portuguese. Finally, this study contributes to both academic and technological advancement by proposing solutions for processing texts in the Portuguese language.

Keywords: STS; NLP; S-BERT; S-BERT.

Lista de figuras

Figura 2.1 Arquitetura Transformers (Vaswani *et al.*,) 19
Figura 2.2 Arquitetura SBERT para calcular pontuações de similaridade 22

Lista de tabelas

Tabela 5.1 Resultados das Previsões e Frases Similares 1 37
Tabela 5.2 Resultados das Previsões e Frases Similares 2 37
Tabela 5.3 Resultados das Previsões e Frases Distintas 1 37
Tabela 5.4 Resultados das Previsões e Frases Distintas 2 38
Tabela 5.5 Resultados das Previsões e Pontuações com maior discrepância 38
Tabela 5.6 Correlação de Pearson ASSIN 39
Tabela 5.7 Correlação de Pearson ASSIN 2 40
Tabela 5.8 Correlação de Pearson SICK-BR 40
Tabela 5.9 Correlação de Spearman para diferentes modelos em ASSIN, ASSIN 2 e SICK-BR 42

Lista de abreviaturas e siglas

ASSIN	Avaliação de Similaridade Semântica e Inferência Textual
ASSIN 2	Segunda Avaliação de Similaridade Semântica e Inferência Textual
BERT	*Bidirecional Encoder Representations from Transformers*
CoSENTLoss	*Contrastive Sentence Loss*
LaBSE	*Language-agnostic BERT Sentence Embedding*
LDA	*Latent Dirichlet Allocation*
MLM	*Masked Language Model*
NLI	*Natural Language Inference*
NSP	*Next Sentence Prediction*
PLN	Processamento de Linguagem Natural
RTE	*Recognizing Textual Entailment*
S-BERT	*Sentence BERT*
SICK-BR	*Sentences Involving Compositional Knowledge - Brazilian Portuguese*
STIL	*Symposium in Information and Human Language Technology*
STS	Similaridade Semântica
TCU	Tribunal de Contas da União
UnB	Universidade de Brasília

Sumário

1 Introdução 13
1.1 Objetivos 14
1.2 Organização do Trabalho 15

2 Fundamentação Teórica 16
2.1 Processamento de Linguagem Natural 16
2.2 Similaridade Semântica 17
2.3 Transformers 18
2.4 BERT 20
2.5 S-BERT 21
2.6 Correlação de Pearson 23
2.7 Correlação de Spearman 24

3 Trabalhos relacionados 26
3.1 ASSIN: Avaliação de Similaridade Semântica e Inferência Textual 26
3.2 A iniciativa ASSIN 2 27
3.3 SICK-BR 28
3.4 Unsupervised Cross-lingual Representation Learning at Scale 28
3.5 Language-agnostic BERT Sentence Embedding 29

4 Metodologias 31
4.1 Pré treinamento 31
4.2 Fine-tuning 32
4.3 Corpora Utilizados 34

5 Resultados 36
5.1 Resultado com o modelo LaBSE 36
5.2 Desempenho dos modelos pré treinados 39
5.3 Resultado comparado com S-BERT 41

6 Conclusão e sugestões 44

Referências 47

1 Introdução

Nunca se produziu tanta informação quanto nos dias atuais. Com tal fato, fica humanamente impossível que todo texto produzido possa ser lido, classificado e analisado minuciosamente por um ser humano. Para esse propósito, há o campo da inteligência artificial chamado Processamento de Linguagem Natural (PLN), que permite que essas tarefas sejam automatizadas.

A Similaridade Semântica Textual (STS) (Fonseca *et al.*, 2016b) é um problema dentro do Processamento de Linguagem Natural (PLN)(Liddy, 2001)que busca medir o grau de semelhança semântica entre textos, um aspecto essencial para tarefas que envolvem a análise de grandes volumes de informações textuais. A importância dessa área é evidenciada por suas inúmeras aplicações práticas, que vão desde a recuperação de informações em sistemas jurídicos até a detecção de plágio, resumo automático de textos e aprimoramento de sistemas de tradução automática. No contexto jurídico, por exemplo, a similaridade textual é amplamente utilizada para identificar precedentes relevantes e organizar jurisprudências, como demonstrado no trabalho de (Costa, 2023), que propõe uma abordagem para medir similaridade entre acórdãos do Tribunal de Contas da União (TCU) para apoiar a formulação de jurisprudências. Este caso reforça a relevância de desenvolver técnicas robustas, eficazes e escaláveis para lidar com os desafios apresentados por domínios específicos e por dados textuais extensos.

Neste trabalho, a motivação principal foi explorar as potencialidades da STS em português, uma língua que, apesar de ser amplamente falada e escrita, ainda apresenta lacunas no desenvolvimento de modelos de linguagem dedicados. Com o advento de modelos de linguagem baseados em redes *Transformer* (Vaswani *et al.*,), como BERT (*Bidirecional Encoder Representations from Transformers*)(Devlin *et al.*,) e suas variantes, surge a oportunidade de se realizar o ajuste fino (*fine-tuning*) específico para a língua portuguesa de tais modelos, utilizando corpora especializados. Essa abordagem permite capturar nuances linguísticas particulares, como morfologia complexa, estruturas sintáticas ricas e expressões idiomáticas, que podem ser mal representadas em modelos multilinguagem. Ao mesmo tempo, modelos como BERT multilinguagem têm demonstrado resultados impressionantes em tarefas variadas, mas ainda precisam ser mais profundamente avaliados em comparação com modelos monolíngues em termos de desempenho em STS.

Outro ponto motivador é a possibilidade de aplicar STS em cenários práticos no Brasil. Além do exemplo jurídico citado, onde a identificação de similaridade entre acórdãos pode otimizar o trabalho de operadores do direito, outros setores podem se beneficiar significativamente dessa tecnologia. Na educação, por exemplo, sistemas de correção automática podem utilizar STS para avaliar a equivalência semântica entre respostas dissertativas de

estudantes e os gabaritos, promovendo uma avaliação mais justa e objetiva. Em plataformas de aprendizado virtual, a STS pode ser usada para recomendar materiais de estudo personalizados com base na semelhança com os conteúdos previamente estudados pelo aluno. No campo da saúde, a similaridade textual pode ser utilizada para identificar diagnósticos médicos similares em prontuários eletrônicos, otimizando a tomada de decisão clínica.

A similaridade textual pode ser usada para identificar duplicações em grandes bases de dados de notícias, auxiliando na organização e recuperação de informações. Além disso, pode ser aplicada para detectar *fake news*, comparando as notícias publicadas com fontes confiáveis e avaliando discrepâncias semânticas.

1.1 Objetivos

Apesar do progresso nos modelos de linguagem, há desafios significativos associados ao uso de STS em português. Primeiramente, a disponibilidade limitada de corpora anotados para tarefas específicas dificulta a criação de modelos de alta qualidade. Trabalhos como o ASSIN(Fonseca *et al.*, 2016a), ASSIN 2(Real; Fonseca; Oliveira,) e SICK-BR(Real *et al.*,), que fornecem *datasets* anotados para tarefas de STS em português, têm sido fundamentais, mas ainda há espaço para expandir e diversificar esses recursos. Em segundo lugar, o desenvolvimento de abordagens que equilibrem precisão e eficiência computacional é essencial para tornar essas soluções viáveis em aplicações práticas em larga escala. Modelos como BERT (Devlin *et al.*,), embora poderosos, são computacionalmente caros, exigindo infraestrutura avançada para treinamento e inferência.

Além disso, uma questão na pesquisa em STS é como abordar a integração de modelos multilinguagem em comparação com modelos monolíngues treinados exclusivamente em português. Modelos multilinguagem, como o xlm-roBERTa (Conneau *et al.*,), oferecem a vantagem de serem treinados em grandes volumes de dados de várias línguas, permitindo que se beneficiem de características comuns entre idiomas. No entanto, sua capacidade de capturar peculiaridades linguísticas pode ser inferior à de modelos treinados exclusivamente em português. Por outro lado, modelos monolíngues apresentam maior especialização, mas podem carecer da robustez obtida a partir de corpora diversificados. Portanto, este trabalho propõe não apenas comparar essas abordagens, mas também investigar combinações ou adaptações que aproveitem o melhor dos dois mundos.

Um fenômeno observado ao se treinar modelos de NLP é a *"Curse of Multilinguality"* (Chang *et al.*,) , a qual consiste na queda de performance ao se adicionar vários idiomas ao modelo: línguas com ampla quantidade de dados perdem performance, enquanto para línguas com menor volume de informações, adicionar mais línguas, principalmente se forem semelhantes, aumenta a performance. Um dos objetivos do trabalho ao se comparar modelos multilinguagens com modelos monolíngues é observar quais são os efeitos ao se aumentar a

quantidade de línguas trabalhadas pelo modelo e como ele dilui esse conhecimento.

Além do avanço tecnológico, este trabalho busca também contribuir para a expansão do conhecimento acadêmico no Brasil. Ao conduzir uma análise sistemática e comparativa de modelos, este trabalho pretende oferecer subsídios que possam orientar futuros trabalhos na área.

Portanto, este trabalho tem como objetivo principal explorar a eficiência de diferentes abordagens para tarefas de STS em português, comparando modelos multilinguagem e monolíngues, com ênfase na adaptação para o português e aplicação prática. Com isso, espera-se contribuir para o avanço de estudos em STS em português.

1.2 Organização do Trabalho

O trabalho está organizado da seguinte forma: no Capítulo 2 - Fundamentação Teórica, são apresentados os principais conceitos e ferramentas utilizados no estudo. O Capítulo 3 - Trabalhos Relacionados contém uma breve revisão literária sobre STS, abordando artigos sobre corpora e modelos empregados. No Capítulo 4 - Metodologias, são descritos o desenvolvimento do código do estudo e os processos realizados. O Capítulo 5 - Resultados apresenta o desempenho dos modelos, exemplos de classificação de pares de frases e comparações com o S-BERT (Reimers; Gurevych, b). Por fim, no Capítulo 6 - Conclusão e Sugestões, é feita uma reflexão sobre o alcance dos objetivos do estudo, além da proposição de possíveis pesquisas futuras na área.

2 Fundamentação Teórica

Para uma melhor compreensão deste estudo, torna-se necessária a introdução e contextualização dos principais conceitos e tecnologias utilizadas. Este capítulo tem como objetivo sintetizar esses tópicos, proporcionando ao leitor uma visão das bases teóricas e metodológicas que fundamentam a pesquisa. Inicialmente, apresenta-se o Processamento de Linguagem Natural (NLP) que é uma subárea da Inteligência Artificial. Em seguida, discute-se o problema Similaridade Semântica Textual, destacando sua importância e aplicações em diversos domínios.

No que se refere às tecnologias empregadas, serão introduzidos os principais modelos que servem de base para os experimentos realizados, como a arquitetura de *Transformers* (Vaswani *et al.*,), que teve como grande contribuição ao PLN permitir o processamento eficiente de sequências textuais de forma paralela. Além disso, são abordados os modelos BERT (Devlin *et al.*,), que tem representações contextuais bidirecionais para o texto, e sua variante S-BERT (*Sentence*-BERT) (Reimers; Gurevych, b), que otimiza a comparação entre sentenças fornecendo *embeddings* de alta qualidade. A compreensão desses modelos é fundamental para que o leitor entenda como foram aplicados no estudo e suas respectivas contribuições para a tarefa de similaridade semântica.

Por fim, são apresentadas as métricas de avaliação utilizadas para quantificar o desempenho dos modelos no contexto da pesquisa. Métricas como a correlação de Spearman e correlação de Pearson são exploradas, destacando-se sua importância para medir a precisão das previsões.

2.1 Processamento de Linguagem Natural

O NLP(Nadkarni; Ohno-Machado; Chapman, 2011) refere-se ao conjunto de técnicas e modelos que permitem que sistemas computacionais processem, analisem e interajam com textos e discursos humanos. Ele engloba uma ampla gama de tarefas, desde simples análises de frequência de palavras até tarefas complexas, como tradução automática, geração de linguagem natural e análise de sentimentos. Modelos modernos de NLP, como o (Mao *et al.*, 2023) e o BERT(Devlin *et al.*,), são baseados em aprendizado profundo, utilizando grandes quantidades de dados para treinar redes neurais capazes de capturar nuances semânticas e contextuais em textos.

O NLP abrange uma ampla gama de tarefas que permitem a compreensão e manipulação de textos e fala pela inteligência artificial.Tais tecnologias já se fazem presentes no cotidiano das pessoas, entre as principais tarefas realizadas estão: análise sintática, que iden-

tifica a estrutura gramatical das sentenças; análise semântica, que interpreta o significado das palavras e frases em seu contexto; e reconhecimento de entidades nomeadas, utilizado para identificar nomes de pessoas, lugares e organizações em textos. Outras aplicações incluem análise de sentimentos, essencial para avaliar opiniões em redes sociais e revisões; tradução automática, que possibilita a comunicação entre diferentes idiomas; geração de texto, usada em *chatbots* e sistemas de criação automática de conteúdos; e reconhecimento de fala e conversão de fala em texto, fundamentais para assistentes virtuais e transcrição de áudio. Além disso, o NLP é aplicado em tarefas como extração de informação, para coletar dados relevantes em grandes volumes de texto; sumarização de texto, que cria resumos concisos de documentos extensos; e detecção de fraude e *spam*, essencial para a segurança digital e filtragem de informações indesejadas.

2.2 Similaridade Semântica

A Similaridade Semântica (Fonseca *et al.*, 2016b) é um problema em Processamento de Linguagem Natural, dedicado a quantificar o grau de proximidade no significado entre palavras, frases ou textos. Essa análise vai além da correspondência literal, capturando nuances semânticas que refletem o contexto e o significado subjacente das expressões linguísticas.

Para que a Similaridade Semântica possa ser calculada, é necessário converter as representações linguísticas utilizadas por humanos em estruturas matemáticas interpretáveis por computadores. Essa conversão geralmente é feita por meio de vetores multidimensionais, que traduzem o conteúdo textual em formatos numéricos adequados para cálculos matemáticos.

Entre as técnicas mais utilizadas para essa conversão estão o *Bag-of-Words* (BoW), o *Term Frequency-Inverse Document Frequency* (TF-IDF) e os *embeddings*. No presente trabalho, optou-se pelo uso de *embeddings* devido à sua capacidade de gerar representações mais ricas e contextuais, especialmente em modelos como o BERT e o S-BERT, que são baseados em redes neurais profundas.

A técnica de *embeddings* inicia-se com um pré-processamento do texto, no qual as palavras são divididas em unidades menores chamadas tokens. Cada *token* é, então, transformado em uma representação numérica e mapeado,em geral por meio de uma Rede Neural Artificial, para um vetor de alta dimensão, que encapsula seu significado dentro do contexto da frase. Essa abordagem é particularmente vantajosa porque considera o ambiente semântico em que cada palavra está inserida, oferecendo uma análise mais precisa e detalhada.

Após a vetorização, é necessário calcular a similaridade entre os vetores resultantes para determinar a proximidade semântica. Os métodos mais comuns incluem:

- **Similaridade por cosseno:** Mede o ângulo entre os vetores e é amplamente utilizada para identificar similaridades semânticas, sendo especialmente eficaz para detectar

relações contextuais mais complexas. Essa técnica foi aplicada no presente trabalho, dado seu desempenho robusto em tarefas de STS.

- **Distância Euclidiana:** Calcula a distância linear entre os vetores, sendo útil em cenários onde a simplicidade é uma prioridade, embora seja menos adequada para capturar nuances contextuais.

O resultado dessa análise é um valor numérico que representa o grau de similaridade semântica entre as palavras, frases ou textos analisados. Este valor é fundamental para diversas aplicações, como recuperação de informações, detecção de parafraseamento, organização de documentos e sistemas de recomendação.

Assim, a Similaridade Semântica combina avanços teóricos e tecnológicos, permitindo uma interpretação mais profunda e contextual da linguagem humana por sistemas computacionais. Este trabalho se alinha a esse contexto, explorando técnicas avançadas de *embeddings* e métricas de similaridade para contribuir na análise semântica em língua portuguesa.

2.3 Transformers

A arquitetura *Transformers*, apresentada no trabalho seminal (Vaswani *et al.*,), marcou um progresso no campo do PLN. Baseada exclusivamente em mecanismos de atenção, essa arquitetura eliminou a necessidade de camadas recorrentes e convolucionais, que até então dominavam os modelos de aprendizado profundo. Essa mudança permitiu maior eficiência computacional, graças à capacidade de paralelizar operações, e um processamento mais rápido para lidar com grandes volumes de dados textuais.

O mecanismo de atenção é o núcleo dos *Transformers*, funcionando ao atribuir pesos diferentes a todas as palavras da entrada, considerando a relação entre elas, independentemente de sua posição na sequência. Isso possibilita uma interpretação mais precisa do contexto global da frase, o que é de extrema importância para tarefas como tradução automática, sumarização de texto e inferência textual. Por exemplo, em uma sentença complexa, o modelo é capaz de identificar quais palavras ou expressões são mais relevantes para determinar o significado geral, evitando que informações importantes sejam ignoradas.

A arquitetura dos *Transformers* é composta por dois blocos principais *encoder* e *decoder*, como mostrado na Figura 2.1, entretanto não é obrigatório o uso de ambos, o BERT por exemplo usa apenas o encoder:

- **Encoder:** Responsável por processar a entrada textual, gerando representações ricas e contextuais por meio de múltiplas camadas de atenção e redes neurais. O encoder é projetado para capturar dependências locais e globais entre os elementos da entrada, garantindo uma compreensão profunda do conteúdo.
- **Decoder:** Utiliza as representações geradas pelo encoder para produzir a saída, seja uma tradução, um texto resumido ou qualquer outra tarefa de geração textual. O

decoder combina informações anteriores com as representações fornecidas pelo *encoder*, permitindo a produção de resultados consistentes e contextualizados.

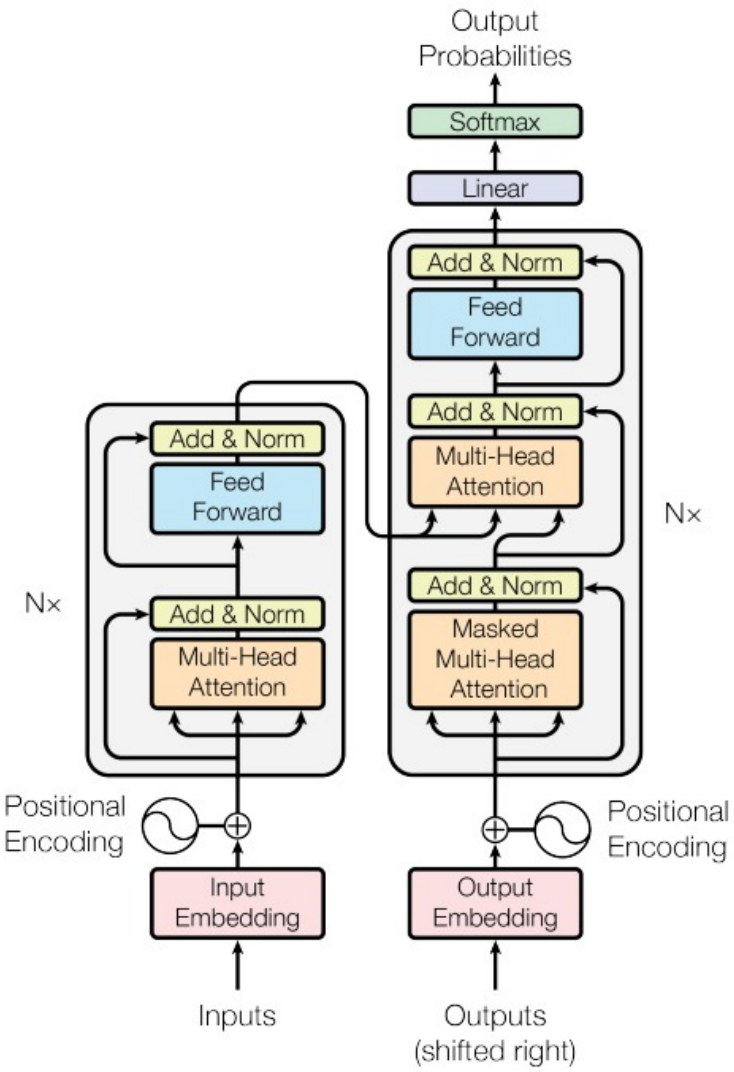


Figura 2.1 – Arquitetura Transformers (Vaswani *et al.*,)

No contexto da Similaridade Semântica Textual (STS), os *Transformers* desempenham um papel crucial ao fornecer representações superiores de sentenças. Modelos baseados em Transformers, como o BERT, utilizam o encoder para criar *embeddings* que capturam as nuances semânticas e contextuais das sentenças de forma bidirecional. Isso significa que cada palavra em uma sentença é representada considerando tanto as palavras que a precedem quanto as que a sucedem, resultando em uma compreensão mais precisa do significado.

Além disso, variantes como o S-BERT (Reimers; Gurevych, b) aprimoram ainda mais a aplicação dos Transformers em STS, ajustando os *embeddings* gerados para tarefas específicas de similaridade. Essa abordagem permite comparar sentenças de maneira eficiente, utilizando métricas como a similaridade por cosseno para medir o grau de proximidade

semântica entre elas.

Os avanços trazidos pelos Transformers mudaram não apenas o panorama do PLN, mas também o estado da arte em STS, oferecendo soluções mais rápidas, precisas e escaláveis para a análise de relações semânticas em textos.

2.4 BERT

O BERT *(Bidirectional Encoder Representations from Transformers)* é um modelo de representação de linguagem desenvolvido pelo Google (Reimers; Gurevych, a) que teve diversas contribuições para o PLN. Baseado na arquitetura *Transformer*, o BERT utiliza mecanismos de atenção para processar sequências de texto de forma altamente eficiente e precisa. Sua característica distintiva é a capacidade de gerar representações contextuais bidirecionais profundas, analisando simultaneamente o contexto à esquerda e à direita de cada palavra. Essa abordagem permite uma compreensão mais rica e detalhada do significado das palavras no contexto global da sentença, superando limitações de modelos unidirecionais.

O BERT é construído exclusivamente sobre o bloco de *encoder* dos *Transformers*, que usa o mecanismo de atenção para atribuir pesos às palavras em uma sentença, capturando as relações semânticas entre elas. Isso permite ao modelo considerar não apenas palavras próximas, mas também aquelas mais distantes na sequência, oferecendo uma visão abrangente do contexto linguístico.

Uma das abordagens do BERT é o pré-treinamento e ajuste fino *(fine-tuning)*. No pré-treinamento, o modelo é exposto a grandes volumes de texto não rotulado, onde aprende padrões linguísticos gerais por meio de duas tarefas principais.A primeira das tarefas é Modelagem de Palavras Mascaradas *(Masked Language Modeling - MLM)*em que o BERT mascara aleatoriamente algumas palavras na entrada e tenta prever essas palavras com base no contexto. Isso força o modelo a entender as relações contextuais entre as palavras restantes.E a outra tarefa é a Predição da Próxima Sentença *(Next Sentence Prediction - NSP)* em que o modelo aprende a prever se uma sentença segue logicamente a outra, o que é crucial para tarefas como inferência textual e compreensão de relacionamentos entre sentenças.

Após o pré-treinamento, o BERT pode ser ajustado para tarefas específicas com um número reduzido de exemplos rotulados. Esse processo é eficiente porque o modelo já possui um entendimento profundo da linguagem, precisando apenas de ajustes para otimizar seu desempenho em aplicações específicas.

O BERT demonstrou um desempenho notável em uma ampla gama de tarefas de PLN, incluindo pergunta e resposta, como em sistemas que extraem informações relevantes de textos com base em perguntas do usuário; inferência textual, que consiste em determinar se uma sentença apoia, contradiz ou é neutra em relação a outra; e análise de sentimentos, que busca identificar emoções ou opiniões expressas em textos.

No campo da Similaridade Semântica Textual, o BERT trouxe avanços significativos ao possibilitar representações contextuais de alta qualidade para sentenças. Sua bidirecionalidade permite capturar nuances semânticas complexas e contextos amplos, essenciais para medir com precisão o grau de similaridade entre textos. O BERT é especialmente eficaz na tarefa de STS porque gera *embeddings* que representam não apenas as palavras individualmente, mas também como elas interagem dentro de uma sentença. Essa capacidade foi amplificada por modelos derivados, como o S-BERT, que otimizam o BERT para calcular similaridades de maneira mais eficiente e escalável. Aplicações de STS com BERT incluem recuperação de informações, detecção de parafraseamento, sistemas de recomendação e análise de documentos jurídicos, áreas onde a compreensão precisa de significados é crucial.

A combinação de sua arquitetura baseada em *Transformers*, bidirecionalidade e o uso de pré-treinamento em grandes corpora posicionou o BERT como um dos modelos no estado da arte no campo. Ele não apenas elevou o desempenho em *benchmarks* padrão de PLN, mas também influenciou o design de modelos subsequentes, como o S-BERT.

2.5 S-BERT

O Sentence-BERT (S-BERT)(Reimers; Gurevych, b) é uma extensão do modelo BERT projetada especificamente para gerar representações otimizadas de sentenças, conhecidas como *embeddings*, que podem ser comparadas de forma eficiente em tarefas como similaridade semântica e agrupamento (*clustering*).

Enquanto o BERT tradicional requer que pares de sentenças sejam processados simultaneamente para calcular a similaridade, o que resulta em uma complexidade computacional elevada, o S-BERT resolve esse problema introduzindo uma arquitetura baseada em redes siamesas e *triplet networks*. Essas arquiteturas permitem que cada sentença seja transformada em um vetor fixo de alta dimensão, o que possibilita a comparação direta de sentenças por meio de medidas como a similaridade por cosseno assim como mostrado na Figura 2.2. Essa modificação reduz o custo computacional, tornando tarefas como busca semântica e clustering viáveis em grandes bases de dados.

O S-BERT utiliza o poder da arquitetura *Transformer* herdado do BERT, mas com um ajuste no treinamento para focar na maximização da similaridade semântica entre pares de sentenças. Durante o treinamento, o modelo é ajustado usando dados anotados, como aqueles provenientes de tarefas de inferência textual natural (NLI), que fornecem pares de sentenças classificadas como equivalentes, contraditórias ou neutras. Isso permite que o S-BERT aprenda a mapear sentenças semanticamente semelhantes para vetores próximos no espaço vetorial.

Além disso, o S-BERT mantém a capacidade do BERT de capturar contextos complexos por meio de sua atenção bidirecional, mas aplica uma etapa adicional de *pooling* (média ou

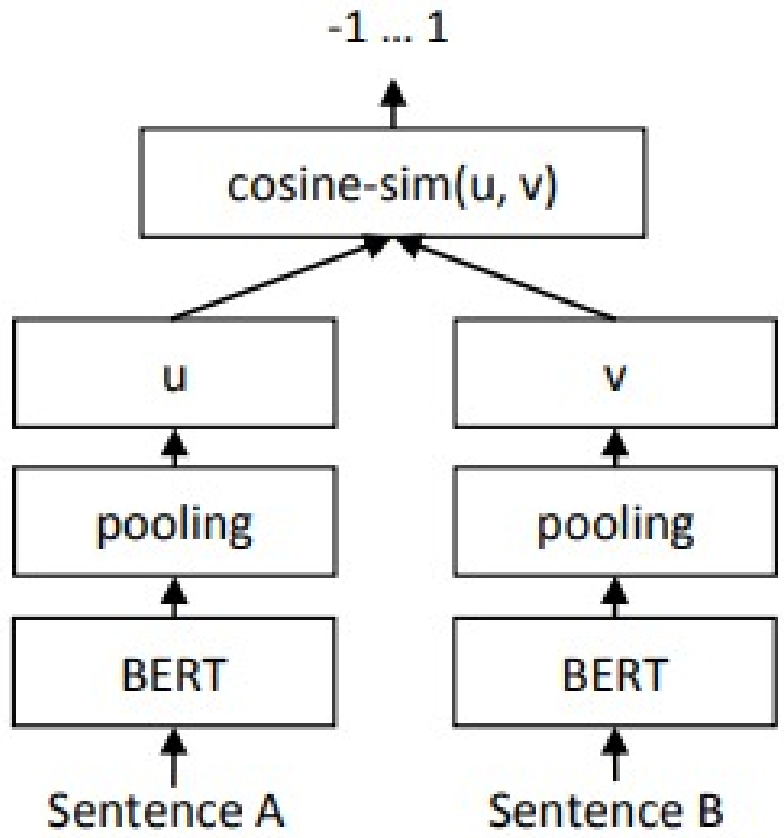


Figura 2.2 – Arquitetura SBERT para calcular pontuações de similaridade

máxima) nos vetores de saída para gerar representações fixas de sentenças. Isso não apenas melhora a precisão em tarefas como similaridade textual, mas também torna o modelo significativamente mais rápido em cenários com grandes volumes de dados.

Na arquitetura do S-BERT, é possível importar diferentes modelos pré-treinados do BERT para atuar como a rede neural siamesa, na qual o S-BERT aprimora a saída do BERT por meio de uma operação de *pooling*, resultando em *embeddings* de sentenças de tamanho fixo. A escolha do modelo BERT pré-treinado a ser utilizado influencia diretamente o desempenho do S-BERT, pois modelos treinados em corpora e idiomas distintos apresentam desempenhos variados. Essa abordagem proporciona ao S-BERT flexibilidade, tornando-o adaptável a diferentes cenários. Neste estudo, essa característica é explorada por meio da avaliação do desempenho de diferentes modelos BERT pré-treinados em corpora da língua portuguesa.

O S-BERT apresenta um bom desempenho em diversas tarefas, como busca semântica, que consiste na identificação eficiente das sentenças mais relevantes em bases de dados extensas; clustering de sentenças, que envolve o agrupamento de sentenças com base em seu significado, sendo útil para a organização de informações; classificação de sentenças,

que se refere à determinação de categorias com base no conteúdo semântico; e resposta a perguntas, que permite a identificação de respostas que correspondem semanticamente às perguntas formuladas.

2.6 Correlação de Pearson

Para avaliar a eficácia dos modelos treinados na tarefa de STS, foi utilizado o coeficiente de correlação de Pearson(Wikipédia, a enciclopédia livre, 2025a) como métrica de desempenho. Essa medida estatística quantifica o grau de alinhamento linear entre as os rótulos atribuídos pelo corpus (valores reais) e as rótulos previstos pelo modelo (valores preditos).

O coeficiente de Pearson, que varia entre -1 e 1, é particularmente adequado para STS, pois permite avaliar a proximidade entre os valores preditos e as expectativas definidas no conjunto de dados de teste. Um valor de correlação próximo de 1 indica que o modelo consegue prever as similaridades semânticas de maneira altamente consistente com os rótulos verdadeiros, demonstrando um forte alinhamento. Por outro lado, valores próximos de 0 indicam pouca ou nenhuma relação linear, sugerindo que o modelo não conseguiu capturar adequadamente os padrões semânticos do corpus.

A aplicação dessa métrica em STS é especialmente relevante devido à natureza contínua dos rótulos utilizadas em muitos benchmarks, onde os valores representam graus de similaridade entre pares de sentenças. Ao comparar diretamente os valores previstos pelo modelo com os rótulos verdadeiros, o coeficiente de Pearson oferece uma visão clara da precisão do modelo na tarefa de medir similaridade semântica.

Além disso, o coeficiente de Pearson não apenas avalia o desempenho geral do modelo, mas também pode ser usado para comparar diferentes configurações de modelos ou estratégias de treinamento, auxiliando na seleção de abordagens mais eficazes.

A fórmula calcula a correlação de Pearson entre os valores de duas variáveis X e Y é:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

Onde:

- x_i: O valor i-ésimo da variável X.
- y_i: O valor i-ésimo da variável Y.
- $\bar{x}$: A média dos valores da variável X, calculada como $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- $\bar{y}$: A média dos valores da variável Y, calculada como $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- n: O número total de observações.

O coeficiente de correlação de Pearson (r) mede a força e a direção da relação linear entre duas variáveis:

- $r = 1$: Correlação linear positiva perfeita (quando uma variável aumenta, a outra também aumenta proporcionalmente).
- $r = -1$: Correlação linear negativa perfeita (quando uma variável aumenta, a outra diminui proporcionalmente).
- $r = 0$: Não há relação linear entre as variáveis (mas podem existir outras formas de dependência).

2.7 Correlação de Spearman

O coeficiente de correlação de Spearman(Wikipédia, a enciclopédia livre, 2025b) é uma medida estatística que assim como o coeficiente de correlação de Pearson, avalia a relação entre duas variáveis. Porém, diferentemente medida vista na seção anterior,aqual avalia relações lineares, esta avalia relações monotônicas, sejam elas lineares ou não. A correlação de Spearman (ρ) mede a força e a direção de uma relação monotônica entre duas variáveis. Seu valor varia entre -1 e 1, e os intervalos indicam o seguinte:

- $\rho = 1$: Relação monotônica positiva perfeita. À medida que uma variável aumenta, a outra também aumenta consistentemente.
- $\rho = -1$: Relação monotônica negativa perfeita. À medida que uma variável aumenta, a outra diminui consistentemente.
- $\rho = 0$: Ausência de relação monotônica. Não há um padrão claro de aumento ou diminuição entre as variáveis.

A correlação de Spearman utiliza os rankings dos dados para calcular a força e a direção da relação entre duas variáveis, sem se preocupar com os valores absolutos, mas sim com a ordem relativa dos dados.Os rankings são classificações atribuídas aos valores de um conjunto de dados com base na sua ordem, normalmente crescente ou decrescente, indicando a posição relativa de cada valor em relação aos demais. Já os empates nos rankings ocorrem quando dois ou mais valores são iguais, ocupando a mesma posição no conjunto. Nesses casos, a solução comum é atribuir a média das posições que os valores empatados ocupariam.

A correlação de Spearman possui a Fórmula Simples Equação (2.2) que é adequada quando não há empates, sendo computacionalmente mais eficiente, mas pode fornecer resultados imprecisos caso existam valores repetidos e a Fórmula Geral Equação(2.3) que calcula a correlação com base nos ranks ajustados, deve ser utilizada quando há empates, garantindo maior precisão em dados reais, embora seja mais complexa e exigente em termos computacionais.

A fórmula válida para conjuntos de dados sem empates ou Fórmula Simples é:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.2}$$

onde:

- d_i é a diferença entre os rankings das duas variáveis para o mesmo item ($d_i = R(x_i) - R(y_i)$).
- n é o número total de observações.

Enquanto que a fórmula baseada no coeficiente de correlação de Pearson aplicado aos ranks ajustados ou Fórmula Geral é:

$$\rho = \frac{\sum(R(x_i) - \bar{R}(x))(R(y_i) - \bar{R}(y))}{\sqrt{\sum(R(x_i) - \bar{R}(x))^2 \sum(R(y_i) - \bar{R}(y))^2}} \tag{2.3}$$

onde:

- $R(x_i)$ e $R(y_i)$ são os rankings ajustados das variáveis x e y.
- $\bar{R}(x)$ e $\bar{R}(y)$ são as médias dos rankings de x e y, respectivamente.

A correlação de Spearman é utilizada em tarefas de Similaridade Textual Semântica (STS) para medir a relação monotônica entre as predições dos modelos com os rótulos. No artigo (Reimers; Gurevych, b), a medida foi empregada para avaliar a performance dos modelos, permitindo uma análise robusta do alinhamento entre as pontuações estimadas pelos modelos e as pontuações anotadas nos corpora de STS.

Diferentemente da correlação de Pearson, que mede exclusivamente relações lineares e pode ser afetada por valores extremos, a correlação de Spearman é baseada em rankings, tornando-a menos sensível a outliers e mais adequada para dados não lineares. Por isso, ela é útil em tarefas de STS, onde a ordem relativa das similaridades previstas é mais relevante do que os valores absolutos.

3 Trabalhos relacionados

Neste capítulo, serão apresentados trabalhos relevantes para a área de Similaridade Semântica Textual, com ênfase nos estudos publicados que abordam os corpora e modelos pré-treinados utilizados nesta pesquisa. O objetivo é proporcionar ao leitor uma visão abrangente sobre a motivação, metodologia e os principais resultados alcançados em investigações que dialogam e contribuem com o presente estudo.

3.1 ASSIN: Avaliação de Similaridade Semântica e Inferência Textual

A Avaliação de Similaridade Semântica e Inferência Textual (ASSIN)(Fonseca *et al.*, 2016a), foi iniciativa proposta concomitantemente com o PROPOR 2016 (Conferência Internacional sobre o Processamento Computacional da Língua Portuguesa), tendo como subtarefas a similaridade semântica (STS) e a inferência textual.

Essa avaliação trouxe o primeiro corpus para essas subtarefas para a língua portuguesa. Esse corpus foi montado a partir de textos jornalísticos fornecidos pelo Google News (Google News, 2025), o trabalho de avaliar as similaridades entre as sentenças foi feito manualmente, em que cada par foi avaliado por quatro pessoas em cada tarefa.

Na tarefa de similaridade semântica o corpus foi classificado em um intervalo de 1 a 5, em que 1 as sentenças são diametralmente diferentes, 2 as sentenças são referentes a fatos distintos e diferentes entre si, mas são do mesmo assunto, 3 as sentenças são um pouco parecidas entre si e podem falar do mesmo assunto ou não, 4 as sentenças são parecidas mas tem algum componente que as distingue e 5 as sentenças são basicamente iguais com alguma pequena diferença no modo de escrita.

Na tarefa de inferência textual, há uma afirmação T e uma hipótese H e se uma pessoa lê T e conclui que H é verdade, temos uma inferência. Em alguns corpora são adicionados textos sem inferências para contradição, estratégia que não foi utilizada no ASSIN. O corpus de inferência textual foi dividido em três classes Implicação, Paráfrase e Nenhuma das duas.

Para a montagem do corpus foi utilizada a técnica de Modelagem de Tópicos *Latent Dirichlet Allocation* (LDA) com suas variações português brasileiro e europeu, a qual é uma forma de modelagem suas respectivas representações em espaços vetoriais em que se atribui pontuação para os textos com base no quanto são semelhantes.

A montagem do corpus passou por três fases,a primeira utilizaram do LDA para evidenciar pares de sentenças parecidas, a segunda foi uma revisão dos pares encontrados para

garantir que eles eram relacionados como também revisão ortográfica e gramatical e na terceira fase quatro pessoas anotaram os pares e deram sua classificação para eles.

A título de estatística sobre o processo de anotações, 36 anotadores fizeram quantidades distintas de anotações, houve o descarte de 11.3% de pares devido à ausência do mínimo de três dentre quatro julgamentos iguais. Foram anotados 10 mil pares, 5 mil para o português europeu e 5 mil para o português brasileiro.

Para verificar a qualidade do processo foi utilizada a média ponderada das correlações de Pearson entre todos os anotadores, com o peso baseado na quantidade de pares anotados. Para cada anotador, calculamos a correlação entre suas pontuações de similaridade com as dos demais, excluindo sua própria anotação. Houve um score de 0.74, o que significa uma boa concordância entre os anotadores.

Outra métrica de qualidade usada foi o desvio padrão médio, o qual mede a variação nos julgamentos de similaridade dos pares. Ele é obtido pela média dos desvios padrão de todos os pares, onde cada um é calculado como o desvio padrão das quatro pontuações em relação à média do par.O desvio padrão médio obtido foi de 0,49 o que indica pequena divergência entre os anotadores.

O corpus ASSIN é dividido em conjuntos de treinamento e de teste, com três mil pares para treinamento e dois mil para teste, em ambas as variantes brasileira e portuguesa. Foram divididas equipes para montarem seus sistemas que se utilizariam das tarefas, os times participantes receberam o conjunto de teste em março de 2016 e tiveram 8 dias para submeter até três resultados.

3.2 A iniciativa ASSIN 2

O ASSIN 2(Real; Fonseca; Oliveira,) é a sequência da iniciativa ASSIN, o qual propôs as tarefas STS e inferência textual.Ele foi proposto no STIL *(Symposium in Information and Human Language Technology)* em Salvador BA no dia 15/10/2019.O ASSIN 2 foi montado com a finalidade de dar continuidade ao seu predecessor em prover à comunidade brasileira um corpus interessante para as tarefas propostas.

O corpus ASSIN 2, diferentemente do ASSIN, foi dividido em duas categorias para a tarefa de inferência textual: Implicação e Não-Implicação, evitando o uso da categoria de paráfrase. A decisão de não utilizar a paráfrase deve-se ao fato de que, por definição, ela implica reciprocidade, tornando desnecessária uma terceira classificação. O corpus foi anotado por pelo menos quatro falantes nativos de português, resultando em aproximadamente 10.000 pares de sentenças.

Os resultados obtidos com o ASSIN 2 apresentaram um desempenho superior à linha de base, demonstrando uma evolução em relação ao seu predecessor. Além disso, observou-se

um crescente interesse da comunidade brasileira no corpus e no campo do Processamento de Linguagem Natural.

3.3 SICK-BR

O SICK-BR (Real *et al.*,) é um corpus desenvolvido para tarefas de inferência textual em português, criado como uma adaptação do corpus original em inglês SICK(Marelli *et al.*, 2014). Ele foi projetado com o objetivo de simplificar e complementar o trabalho iniciado pelo ASSIN, trazendo uma abordagem mais básica e acessível para a comunidade de Processamento de Linguagem Natural em português. Diferentemente do ASSIN, que contém sentenças mais complexas, o SICK-BR se concentra em fenômenos linguísticos mais simples, alinhando-se à estrutura e propósitos do corpus internacional SICK.

O processo de criação do SICK-BR envolveu várias etapas: a tradução inicial por máquina do corpus SICK para o português, seguida de revisões detalhadas por especialistas linguistas. Essas revisões buscaram preservar as relações de inferência (Implicação, Contradição e Neutro) e as pontuações de similaridade semântica do corpus original, ao mesmo tempo em que mantinham a naturalidade das sentenças no português brasileiro. Para garantir a consistência das traduções, foram criadas diretrizes rigorosas e um glossário para padronizar termos e expressões recorrentes no corpus.

Além disso, o SICK-BR manteve o alinhamento com o corpus original em nível de pares de sentenças, permitindo que os pesquisadores explorem a inferência textual de forma paralela entre os dois idiomas. Essa característica é especialmente útil para estudos de transferência linguística e aprendizado multilingue. No entanto, devido às diferenças linguísticas entre inglês e português, algumas adaptações foram necessárias, como a revisão de traduções que poderiam alterar as relações de inferência ou gerar duplicações no corpus traduzido.

Com aproximadamente 10 mil pares de sentenças anotadas para relações de inferência e similaridade semântica, o SICK-BR representa uma importante contribuição para o avanço da pesquisa em inferência textual em português, fornecendo uma base simplificada, porém robusta, para treinar e testar modelos de PLN.

3.4 Unsupervised Cross-lingual Representation Learning at Scale

O artigo (Conneau *et al.*,) tem como objetivo aprimorar as técnicas de aprendizado de linguagem por meio de Processamento de Linguagem Natural (PLN), apresentando o modelo XLM-RoBERTa (XLM-R), uma abordagem baseada em Transformers multilingues. O XLM-R utiliza o Masked Language Modeling (MLM), uma tarefa de modelagem de

linguagem mascarada, e foi pré-treinado em 100 línguas diferentes, abrangendo uma grande diversidade de idiomas com variados níveis de recursos. Esse modelo foi desenvolvido para melhorar o desempenho em tarefas como classificação entre línguas, atribuição de rótulos para sequências e resposta a questões. O XLM-R representa um avanço na transferência de aprendizado entre línguas, fornecendo um estado da arte na comparação de desempenho entre modelos multilingues e monolíngues.

Um dos principais objetivos do artigo é explorar a ideia do treinamento multilinguagem e os desafios associados ao "curse of multilinguality", ou maldição da multilinguagem. Esse fenômeno ocorre quando o aumento do número de línguas no modelo leva à diluição da capacidade de aprendizado, prejudicando o desempenho nas línguas com menos recursos. No entanto, o XLM-R aborda esse desafio ao ampliar a capacidade do modelo e ao otimizar o uso de dados, permitindo que o modelo se beneficie de múltiplas línguas sem sacrificar a performance por idioma. O modelo foi treinado utilizando dados do CommonCrawl (Wenzek *et al.*,), que oferece uma base de dados abrangente e limpa, essencial para melhorar o desempenho especialmente em línguas de baixo recurso.

Os resultados do XLM-R em comparação com outros modelos como o mBERT e modelos monolíngues mostraram avanços notáveis, especialmente em idiomas de baixo recurso. O XLM-R obteve um aumento de 14.6% de precisão média no benchmark XNLI (Cross-lingual Natural Language Inference), 13% de aumento no F1 score no MLQA (Multilingual Question Answering), e um aumento de 2.4% no F1 score no NER (Named Entity Recognition). Esses ganhos indicam que o XLM-R não apenas supera modelos anteriores, mas também é capaz de realizar uma transferência eficaz entre línguas, mantendo um bom desempenho mesmo nas línguas que tradicionalmente têm menos dados para treinamento. Além disso, o modelo demonstrou a possibilidade de modelagem multilingue sem comprometer a performance individual de cada idioma, apresentando resultados comparáveis aos de modelos monolíngues fortes, como RoBERTa e BERT, no GLUE e XNLI.

Dessa forma, o artigo contribui para o avanço do PLN, demonstrando que, ao aumentar a capacidade do modelo e utilizar grandes volumes de dados multilingues de forma eficiente, é possível mitigar a maldição da multilinguagem e alcançar o estado da arte em tarefas de NLP. Esses resultados destacam a eficácia do treinamento multilingue no aprimoramento da compreensão e transferência de informações entre línguas, especialmente para aquelas com poucos recursos linguísticos.

3.5 Language-agnostic BERT Sentence Embedding

Com o objetivo de explorar abordagens multilingues para a tarefa de Similaridade Semântica Textual (STS), foi proposto o Language-agnostic BERT Sentence Embedding (LaBSE) (Feng *et al.*,), um modelo projetado para gerar *embeddings* de sentenças em 109 idiomas.

Baseado na arquitetura *Transformer*, o LaBSE combina técnicas como *Masked Language Modeling* (MLM) e *Translation Language Modeling* (TLM) para capturar representações ricas e universais tanto para tarefas monolíngues quanto entre diferentes idiomas.

O LaBSE utiliza um modelo Dual Encoder, no qual sentenças de origem e destino são processadas separadamente por codificadores baseados no BERT. Essas representações são comparadas em um espaço compartilhado de *embeddings*.

A principal inovação do LaBSE está em sua capacidade de gerar *embeddings* com boa qualidade em cenários que não se tem ampla disponibilidade de informação sobre a linguagem. Isso foi alcançado por meio de treinamento em um corpus bilíngue construído a partir de páginas da web usando uma mineração com mais de 6 bilhões de pares de sentenças e dados monolíngues do CommonCrawl(Common Crawl, 2025) e da Wikipédia(Wikimedia Foundation, 2025). Além disso, o modelo foi pré-treinado utilizando codificadores iniciais baseados em grandes modelos de linguagem, reduzindo a necessidade de dados paralelos em até 80%.

4 Metodologias

Neste capítulo, serão apresentados os modelos pré-treinados utilizados na pesquisa, o processo de *fine-tuning* realizado e os corpora empregados nesse processo. A seção aborda os modelos baseados na arquitetura BERT escolhidos que serão utilizados pelo S-BERT, discutindo as motivações para a escolha de cada um, além de destacar suas principais características e particularidades que os tornam relevantes para o estudo.

Além disso, será detalhado o processo de *fine-tuning* realizado neste trabalho, incluindo a descrição dos principais parâmetros configurados e a justificativa para suas escolhas. Também serão apresentados os corpora utilizados nesse processo, evidenciando sua importância para os modelos e para a realização de análises mais precisas no contexto de STS em português.

4.1 Pré treinamento

No contexto do pré-treinamento do BERT, duas técnicas foram desenvolvidas para permitir que o modelo aprenda representações profundas e contextuais da linguagem: o *Masked Language Model* (MLM) e o *Next Sentence Prediction* (NSP).

Em MLM, utilizada para implementar a bidirecionalidade na arquitetura do BERT, algumas palavras de entrada são mascaradas aleatoriamente durante o pré-treinamento, e o modelo é treinado para prevê-las com base no contexto fornecido pelas palavras restantes. Isso permite que o BERT aprenda representações contextuais ricas, capturando informações não apenas do texto que precede uma palavra, mas também do que a sucede. Esse aprendizado bidirecional é um diferencial importante em relação a arquiteturas anteriores, que eram unidirecionais.

Por outro lado, para tarefas que exigem o entendimento de relações entre sentenças, como pergunta e resposta ou inferência textual, foi introduzida a técnica de NSP. Durante o pré-treinamento, o BERT é exposto a pares de sentenças A e B, sendo que, em 50% dos casos, a sentença B realmente sucede a sentença A, e nos outros 50%, B é uma sentença aleatória não relacionada a A. Essa abordagem permite que o modelo aprenda a identificar se duas sentenças estão semanticamente conectadas.

Essas duas técnicas de pré-treinamento são essenciais para a arquitetura do BERT, tornando-o eficaz em problemas de classificação de texto, similaridade semântica e sistemas de perguntas e respostas.

Para diminuir os custos de treinamento, são utilizados modelos de linguagem pré-treinados, que consistem em modelos de *deep learning* que passaram por treinamento,

utilizando as técnicas de NSP, MLM, entre outras, previamente em corpora massivos. É interessante utilizá-los para economizar recursos em comparação a treinar seu próprio modelo do zero, visto que isso demandaria tempo e muitos recursos computacionais.

No presente trabalho foram utilizados 4 modelos pré treinados, que foram o XLM-RoBERTa, LaBSE, MiniLM-L12-v2 e o bert-base-portuguese. O modelo XLM-RoBERTa (Conneau *et al.*,) é um modelo multilíngue, treinado para mais de 100 línguas através do corpus filtrado do CommonCraw (Wenzek *et al.*,) utilizando como tarefa de treino o MLM em apenas textos "crus", ou seja, textos que não foram classificados por humanos, seu uso pode ser interessante para comparar com modelos que tiveram interferência humana no treinamento.

O LaBSE (Feng *et al.*,) também é um modelo pré treinado em mais de 100 línguas e que pode ser utilizado para diversas tarefas dentro de PLN. O LaBSE também é baseado no modelo BERT, mas foi ajustado especificamente para produzir *embeddings* de sentenças ao invés de palavras, o que permite ele alinhar sentenças entre diferentes línguas, para melhorar sua perfomance em problemas de tradução.Ele foi pré treinado em diferentes corpora, entre eles o CCMatrix (Schwenk *et al.*, 2020), TED test set(Qi *et al.*, 2018) e Tatoeba(Artetxe; Schwenk, 2019).

O MiniLm[1] (Wang *et al.*, 2020) é pré treinado para mais de 50 línguas, seu nome deriva de *Minimal Language Model*, o que também reflete em seu design, pois ele foi projetado para ser mais leve, eficiente e rápido do que modelos mais robustos como o RoBERTa. Ele foi treinado utilizando uma técnica conhecida como destilação de conhecimento, em que ele aprende com algum modelo mais robusto sobre as representações e tenta replicá-las por meio de sua arquitetura mais simples, assim diminuindo significativamente a demanda computacional.

O bert-base-portuguese[2] é o modelo BERTimbau (Souza; Nogueira; Lotufo, 2020), adicionalmente pré-treinado com o corpus do ASSIN 2, ou seja, pré-treinado na tarefa de STS em português, o que pode ser interessante para o objetivo deste trabalho, ao comparar com abordagens multilinguagens.

4.2 Fine-tuning

O *fine-tuning* consiste em ajustar (fazer uma rodada de treinamento) um modelo pré-treinado para um conjunto de dados e tarefa específicos, aproveitando o conhecimento já adquirido pelo modelo durante o pré-treinamento. Esse processo permite que o modelo seja

[1] Esse modelo pode ser encontrado no HuggingFace como "sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2"

[2] Esse modelo pode ser encontrado no HuggingFace como "ricardo-filho/bert-base-portuguese- cased-nli-assin-2"

adaptado para as nuances dos dados e para os objetivos da tarefa em questão, maximizando seu desempenho. O *fine-tuning* dos modelos baseados no BERT foi configurado com parâmetros otimizados para garantir um treinamento eficiente e controlado. Foram utilizados os seguintes parâmetros:

- **Número de épocas = 10**: Define o número de épocas passagens completas sobre os dados para o treinamento do modelo. São configuradas 10 épocas para treinamento, visto que o conjunto de dados não é extenso, o que permite um número maior de épocas.
- **Learning rate de** 2×10^{-5}:A taxa de aprendizado controla a velocidade com que os pesos do modelo são ajustados. Um valor pequeno, como 2×10^{-5}, é adequado para fine-tuning de modelos pré-treinados, evitando mudanças abruptas nos pesos e promovendo ajustes suaves.
- **Passos de validação = 100**: Configura a frequência de avaliação com base no número de passos de treinamento. A cada 100 passos, o desempenho é avaliado usando o conjunto de validação, possibilitando monitorar a evolução do modelo.

Os parâmetros foram configurados para garantir a viabilidade do treinamento no ambiente do Google Colab (Google, 2025), utilizando uma GPU NVIDIA T4, que oferece recursos adequados para experimentos com modelos baseados no BERT.

No treinamento de modelos de *deep learning*, a utilização de uma função de perda *(loss function)* é essencial para quantificar o desempenho do modelo ao ajustar suas previsões. A função de perda mede a diferença entre o valor previsto pelo modelo e o valor real atribuído a cada par de sentenças, calculando o erro. Esse erro é então usado para atualizar os pesos do modelo, com o objetivo de minimizar a discrepância e melhorar a precisão nas próximas iterações de treinamento.

Para o presente trabalho, foi escolhida a CoSENTLoss (Huang *et al.*, 2024) como função de perda, devido à sua adequação às tarefas de STS. A escolha foi fundamentada nas evidências apresentadas em (Reimers; Gurevych, b), que destacou a eficácia do uso da similaridade por cosseno como métrica principal para avaliar a proximidade entre os *embeddings* das frases. Essa abordagem se alinha diretamente com a tarefa de prever graus contínuos de similaridade, típicos de corpora como ASSIN, ASSIN 2 e SICK-BR.

A CoSENTLoss oferece vantagens específicas para STS:

- **Adaptação para Pontuações Contínuos:** Diferentemente de funções de perda projetadas para trabalhar com valores discretos, a CoSENTLoss é projetada para lidar com valores contínuos, tornando-a adequada para tarefas que envolvem previsões de similaridade em uma escala numérica.
- **Eficiência na Comparação de *Embeddings*:** A CoSENTLoss avalia a proximidade dos *embeddings* de duas frases, utilizando a similaridade por cosseno como critério.

Isso permite uma representação mais fiel das relações semânticas entre as frases.

- **Generalização:** A CoSENTLoss incentiva uma separação clara entre pares de frases similares e não similares, melhorando a capacidade do modelo de generalizar para dados novos.

Durante o treinamento no presente trabalho, a CoSENTLoss foi integrada ao processo de *fine-tuning* dos modelos baseados no BERT. Essa configuração potencializou a adaptação dos modelos aos corpora de STS em português (ASSIN, ASSIN 2 e SICK-BR), aproveitando a robustez do pré-treinamento e refinando as previsões para refletir as nuances semânticas desses conjuntos de dados. A escolha dessa função de perda foi crucial para alcançar melhores resultados na tarefa de STS, permitindo que o modelo aprendesse a capturar relações semânticas com maior precisão e relevância para a língua portuguesa.

Para explorar os modelos pré-treinados selecionados e os corpora de STS disponíveis em português, foi realizado o *fine-tuning* com os corpora ASSIN, ASSIN 2 e SICK-BR. Essa abordagem permitiu avaliar como cada modelo se adapta a diferentes características dos corpora, explorando suas capacidades de generalização e adaptação.

O *fine-tuning* é especialmente importante na tarefa de STS porque permite ao modelo capturar as nuances específicas das relações semânticas entre pares de sentenças no idioma-alvo. Sem o ajuste, modelos pré-treinados podem falhar em refletir adequadamente as características do corpus utilizado, limitando seu desempenho. Por meio do *fine-tuning*, o modelo é capaz de identificar padrões contextuais mais relevantes, aprimorando sua capacidade de prever graus de similaridade com maior precisão e robustez.

4.3 Corpora Utilizados

Nos experimentos realizados neste trabalho, foram utilizados os corpora ASSIN, ASSIN 2 e SICK-BR para avaliar o desempenho de modelos pré-treinados baseados no BERT, antes e após a aplicação do *fine-tuning*.A proporção exata entre os conjuntos pode variar conforme o corpus adotado, mas, em geral, foi mantida a distribuição clássica de aproximadamente 65% para treinamento, 5% para validação e 30% para teste, respeitando o balanceamento das classes de similaridade sempre que possível. A escolha desses corpora foi fundamentada em suas contribuições históricas e técnicas para o avanço das tarefas de STS em português, sendo complementares em suas características.

Cada corpus foi dividido em três conjuntos: treino, validação e teste. Essa separação é essencial para assegurar um treinamento eficaz e uma avaliação confiável dos modelos:

O conjunto de treino contém a maior parte dos pares de sentenças e é utilizado para ajustar os parâmetros do modelo. Durante esta etapa, o modelo aprende a associar os pares de sentenças às suas respectivas anotações, no caso do presente trabalho, pontuações de similaridade.

O conjunto de validação é empregado durante o desenvolvimento do modelo, proporcionando uma estimativa de seu desempenho em dados não vistos diretamente no treinamento. Este conjunto auxilia na escolha de hiperparâmetros e na identificação de problemas como overfitting, garantindo que o modelo mantenha um bom desempenho em cenários mais amplos.

O conjunto de teste é reservado exclusivamente para a avaliação final. Ele simula o desempenho do modelo em condições reais de uso, sendo essencial para fornecer métricas imparciais e robustas de performance.

O ASSIN, SICK-BR e ASSIN 2 são corpora fundamentais para o estudo de STS em português, o ASSIN foi pioneiro na área, sendo desenvolvido para a PROPOR 2016, contendo pares de sentenças jornalísticas. Já o SICK-BR, adaptado do corpus inglês SICK, sendo criado por tradução automática e validação manual, tornando-se um complemento importante para análise de modelos em português. Por fim, o ASSIN 2, a iniciativa mais recente, combina características do ASSIN e do SICK-BR, criando pares de sentenças mais diversos e equilibrados entre simplicidade e complexidade. Sua inclusão neste trabalho permite explorar o impacto de corpora criados e adaptados em português, avaliando o estado da arte e avançando a pesquisa em STS.

5 Resultados

Nesta seção, são apresentados os resultados obtidos após o *fine-tuning* dos modelos, analisando seu desempenho na tarefa de STS. Serão exibidos exemplos de classificação de pares de sentenças, destacando como os modelos interpretaram a similaridade entre os textos. Além disso, será feita uma análise quantitativa do desempenho dos modelos, comparando os resultados antes e depois do *fine-tuning*, utilizando métricas apropriadas para avaliar sua eficácia.

5.1 Resultado com o modelo LaBSE

Nesta seção, são apresentados os resultados obtidos nos experimentos realizados com os modelos pré-treinados, assim como após a aplicação do *fine-tuning* utilizando diferentes corpora. Além disso, serão discutidos exemplos específicos de como os pares de sentenças foram classificados, destacando a performance dos modelos ajustados.

A correlação de Pearson entre as similaridades preditas pelos modelos e as anotações humanas foi calculada com base em aproximadamente 3.000 pares de sentenças do conjunto de teste do corpus ASSIN 2.

A seguir, analisamos as previsões realizadas pelo modelo pré-treinado LaBSE, ajustado com o *fine-tuning* utilizando o corpus ASSIN 2.O modelo classifica os pares de sentenças em um intervalo contínuo [-1,1], o qual difere das anotações originais do corpus, que estão no intervalo [1,5]. Para alinhar as predições do modelo aos valores do corpus, foi aplicada a seguinte transformação linear:

$$\text{Predição} = 2 \times \text{Predição feita pelo modelo} + 3 \tag{5.1}$$

Essa transformação possibilita uma comparação direta entre as predições do modelo e as anotações do corpus, assegurando que os resultados sejam interpretados no mesmo intervalo.

Na tabela 5.1, destacam-se pares de sentenças que foram classificados com graus altos de similaridade, evidenciando a capacidade do modelo ajustado de capturar nuances textuais, como diferenças mínimas de ordem ou variação lexical, sem perder o sentido semântico geral. Esses resultados refletem a eficácia do *fine-tuning* na tarefa de STS, particularmente no contexto de sentenças em português.

Nas Tabelas 5.1 e 5.2, as sentenças analisadas pelo modelo são praticamente idênticas em estrutura e significado, o que resultou na atribuição de uma pontuação alta de similaridade pelo modelo. Esse valor está alinhado ao valor anotado pelo corpus, evidenciando a

Tabela 5.1 – Resultados das Previsões e Frases Similares 1

Predição	Pontuação Anotada
4.9993	5.0000
Frase 1	Uma jovem menina de azul está pulando no ar
Frase 2	Uma menina jovem de azul está pulando no ar

Tabela 5.2 – Resultados das Previsões e Frases Similares 2

Predição	Pontuação Anotada
4.9993	5.0000
Frase 1	A criança jovem está andando de motinho de três rodas pela calçada
Frase 2	Uma criança jovem está andando de motinho de três rodas pela calçada

capacidade do modelo de reconhecer e quantificar adequadamente a proximidade semântica entre sentenças tão similares. Esse comportamento demonstra que o modelo compreende nuances textuais como ordem das palavras e escolha de sinônimos, que, embora não alterem o significado geral, são cruciais para tarefas de STS.

Nos seguintes exemplos, Tabelas 5.3 e 5.4, foram identificadas sentenças com baixa pontuação anotada e pontuação mediana atribuída pelo modelo.

Foi notado que em alguns casos o modelo atribuiu pontuações maiores do que as que os anotadores humanos classificaram com valores baixos. Alguns dos possíveis fatores podem ser: o foco em similaridades superficiais, como palavras ou estruturas gramaticais semelhantes, ignorando diferenças semânticas importantes no contexto global das frases. Como também, o modelo pode ter dificuldade em captar modificadores cruciais ou em identificar contradições explícitas e implícitas, o que resulta em avaliações equivocadas. A dependência excessiva de similaridade lexical, aliada à insensibilidade a diferenças nos sujeitos ou objetos das ações, também contribui para interpretações errôneas. Sentenças que possuem múltiplos significados ou relações semânticas mais complexas, como hierarquias ou conexões sutis, são frequentemente mal interpretadas devido à falta de compreensão profunda ou de conhecimento externo. Por fim, possíveis inconsistências ou ruídos no corpus de treinamento podem levar o modelo a generalizar de forma inadequada.

Tabela 5.3 – Resultados das Previsões e Frases Distintas 1

Predição	Pontuação Anotada
2.7441	1.0000
Frase 1	Uma mulher está rachando ovos
Frase 2	Um homem está falando com o menino

Como mostrado nas Tabelas 5.3 e 5.4 o modelo atribuiu uma pontuação mediana, enquanto o corpus anotou pontuação mínima, apontando baixa similaridade. O modelo provavelmente focou no sujeito da frase ser um humano, ignorando as diferenças nos sujeitos que apesar de ambos serem humanos, são distintos ("homem"e "mulher") e nos contextos. Isso pode refletir uma leve dificuldade do modelo em compreender diferenças de contexto mais sutis.

Tabela 5.4 – Resultados das Previsões e Frases Distintas 2

Predição	**Pontuação Anotada**
2.8266	1.0000
Frase 1	Uma mulher está quebrando ovos
Frase 2	Um homem sem camisa está posando com uma menina que está caminhando em um parque

A discrepância entre as pontuações mostradas na Tabela 5.5 possivelmente se deve ao fato de que o modelo, ao avaliar a semelhança semântica, pode ter se concentrado em aspectos superficiais das frases, como a cor citada nas frases e ações realizadas por ambos sujeitos serem a mesma. O modelo pode ter interpretado esses elementos em comum como indicadores de uma maior semelhança, ignorando as grandes diferenças no contexto.

Tabela 5.5 – Resultados das Previsões e Pontuações com maior discrepância

Predição	**Pontuação Anotada**
4.5460	1.5000
Frase 1	Um cachorro preto e um branco estão correndo alegremente na grama
Frase 2	Uma pessoa negra vestindo branco está correndo alegremente com o cachorro na grama

Existem várias razões potenciais para essa discrepância entre as predições do modelo e os valores mais baixos atribuídos pelos anotadores humanos. Primeiramente, o modelo pode ter sido excessivamente influenciado por padrões gerais e associações amplas que são comuns em muitas frases, como a presença de termos que se referem a ações físicas, a agentes humanos ou a contextos de interação social. Mesmo quando as frases não têm uma relação direta, o modelo pode identificar uma sobreposição superficial de características, levando-o a atribuir um valor mais alto do que seria adequado.

Além disso, o modelo pode ter dificuldades em distinguir diferenças mais sutis no significado das frases, especialmente em casos em que as sentenças não apresentam relações evidentes de semelhança, como em pares que falam sobre contextos muito diferentes ou que não compartilham elementos semânticos óbvios. Isso pode ocorrer porque o modelo, apesar de ser treinado com grandes volumes de dados, pode não ser capaz de captar a complexidade

das nuances semânticas em frases com baixa similaridade, especialmente quando essas diferenças são mais contextuais ou subjetivas, algo que os anotadores humanos tendem a perceber com maior precisão.

5.2 Desempenho dos modelos pré treinados

Como já mencionado, no contexto de STS, a correlação de Pearson ajuda a verificar se o modelo é capaz de prever pontuações que seguem o mesmo padrão das anotadas no corpus, mesmo que não sejam exatamente iguais. Um coeficiente de correlação alto sugere que o modelo está capturando bem as relações semânticas entre as sentenças, enquanto valores baixos indicam que as previsões podem estar desalinhadas com as anotações humanas.

Essa análise fornece uma métrica quantitativa clara para comparar o desempenho do modelo antes e após o fine-tuning, bem como avaliar o impacto de diferentes corpora e configurações de treinamento.

Tabela 5.6 – Correlação de Pearson ASSIN

Modelos	**Sem Fine Tuning**	**Com Fine Tuning**
xlm-roberta-base	0.4091	0.7248
LaBSE	0.7286	0.8098
MiniLM-L12-v2	0.6853	0.7798
bert-base-portuguese	0.6098	0.7720

Ao realizar a análise de desempenho dos modelos pré-treinados na tarefa de STS utilizando o corpus ASSIN, a métrica de correlação de Pearson foi empregada para medir o alinhamento entre as pontuações previstas pelos modelos e aquelas anotadas no corpus. Os resultados mostrados na Tabela 5.6 evidenciam a dificuldade do modelo multilinguagem XLM-RoBERTa-base em comparação com outros modelos também multilinguagem, como o LaBSE e o MiniLM-L12-v2, e também uma performance pior do que o modelo em português bert-base-portuguese.

Além disso, observa-se um aumento de performance após a aplicação do *fine-tuning*. Esse aumento reflete a capacidade dos modelos pré-treinados de se adaptarem às nuances do português, especialmente quando ajustados com dados específicos da tarefa.

Ao avaliar na Tabela 5.7 o desempenho dos modelos pré-treinados no conjunto de teste do ASSIN 2 sem a aplicação do *fine-tuning*, observa-se uma queda de performance em relação ao corpus ASSIN, seu antecessor, nos modelos xlm-roberta-base, LaBSE e bert-base-portuguese . Essa redução pode ser atribuída ao fato de que o ASSIN 2 é mais detalhado e diverso, incorporando transformações lexicais e sintáticas que ampliam sua complexidade. Modelos sem *fine-tuning*, portanto, podem não ser capazes de capturar as nuances semânticas

Tabela 5.7 – Correlação de Pearson ASSIN 2

Modelos	Sem Fine Tuning	Com Fine Tuning
xlm-roberta-base	0.3567	0.7704
LaBSE	0.6972	0.8232
MiniLM-L12-v2	0.7721	0.8385
bert-base-portuguese	0.4099	0.8039

mais refinadas presentes no ASSIN 2, limitando seu desempenho.

Entretanto, ao realizar o *fine-tuning* nos modelos utilizando o corpus ASSIN 2, os resultados tornam-se mais expressivos do que aqueles observados no ASSIN. Indicando que o ASSIN 2 conseguiu melhorias na tarefa de STS em português.

Um aspecto intrigante na análise foi a queda de desempenho do modelo bert-base-portuguese no ASSIN 2, quando comparado a outros modelos multilinguagem, como o LaBSE e o MiniLM-L12-v2, no cenário sem *fine-tuning*. Considerando que o bert-base-portuguese foi pré-treinado com dados do ASSIN 2, era esperado que ele apresentasse um desempenho superior nesses dados. No entanto, a performance inferior sugere que, mesmo com o pré-treinamento, o modelo não foi capaz de capturar plenamente as características específicas do corpus durante o ajuste inicial. Isso levanta questões importantes sobre os limites do pré-treinamento em capturar nuances semânticas sem um *fine-tuning* específico, especialmente em tarefas como STS, onde o contexto e a relação entre sentenças desempenham papéis cruciais.

Tabela 5.8 – Correlação de Pearson SICK-BR

Modelos	Sem Fine Tuning	Com Fine Tuning
xlm-roberta-base	0.4208	0.8460
LaBSE	0.7247	0.8773
MiniLM-L12-v2	0.8037	0.8859
bert-base-portuguese	0.3717	0.8673

Os resultados mostrados na Tabela 5.8 mostram que o corpus SICK-BR, após a realização do *fine-tuning*, apresentou um desempenho levemente superior ao ASSIN 2, destacando o sucesso na tradução e adaptação do corpus original SICK para o português. Esse resultado reforça a robustez do processo de tradução e validação adotado, bem como a capacidade do SICK-BR de capturar relações semânticas de maneira consistente e alinhada às expectativas humanas, mesmo em um idioma distinto do original.

Além disso, observa-se que os modelos pré-treinados que obtiveram o melhor desempenho no SICK-BR foram os modelos LaBSE e MiniLM, ambos de arquitetura multilinguagem, superando o bert-base-portuguese, que foi especificamente treinado com corpora em por-

tuguês. Esse fenômeno pode se dar pelo volume de dados no pré-treinamento, visto que o LaBSE foi pré-treinado em um conjunto massivo de dados multilinguagem, abrangendo 109 idiomas, o que permite uma maior generalização e robustez em diferentes contextos semânticos, mesmo em português. Esse amplo treinamento pode contribuir para a habilidade do modelo em lidar com nuances complexas, como as presentes no SICK-BR. Como também, pelo fato do MiniLM utilizar técnicas de destilação de conhecimento, onde aprende a partir de um modelo maior e mais robusto que também foi treinado com um volume de dados massivo, possivelmente contribuiu para o desempenho positivo do modelo.

O desempenho inferior do bert-base-portuguese, mesmo sendo um modelo monolíngue treinado especificamente para o português, pode ser atribuído à menor quantidade de dados utilizados durante seu pré-treinamento em comparação com os outros modelos. Apesar de focar exclusivamente no idioma, a falta de um conjunto de treinamento tão amplo quanto o LaBSE e o uso de métodos avançados, como destilação no caso do MiniLM, talvez esteja limitando sua capacidade de generalização.

Esses resultados ressaltam a relevância de métodos de pré-treinamento escaláveis e o impacto de técnicas como a destilação de conhecimento em tarefas desafiadoras de PLN como a Similaridade Semântica Textual. Modelos que incorporam grandes volumes de dados multilinguagem, como o LaBSE, ou utilizam estratégias eficientes de compressão, como o MiniLM com destilação de conhecimento, exibiram maior capacidade de generalização e robustez, mesmo em cenários específicos como o português.

Além disso, os resultados evidenciam o valor dos corpora SICK-BR e ASSIN 2 como *benchmarks* eficazes para o idioma português. O SICK-BR, derivado de uma tradução bem-sucedida do corpus original em inglês, e o ASSIN 2, com suas transformações lexicais e sintáticas aprimoradas. Ambos os corpora contribuem para o avanço do estado da arte em STS no contexto do português.

Outro ponto que se destaca é a importância do *fine-tuning* no ajuste dos modelos para tarefas específicas. Em todos os corpora analisados, os resultados demonstraram melhorias no desempenho após o fine-tuning, indicando que essa etapa é eficaz para adaptar os modelos pré-treinados às características únicas dos dados e das tarefas.

5.3 Resultado comparado com S-BERT

No trabalho que propõe o modelo Sentence-BERT (SBERT) (Reimers; Gurevych, b), o seu desempenho do modelo em diversas tarefas de STS foi avaliado por meio da correlação de Spearman multiplicada por 100, uma métrica padrão nesse domínio. Os resultados demonstraram que o modelo SRoBERTa-NLI-large obteve a maior média nos *benchmarks* avaliados, alcançando 76,68. Este resultado foi obtido a partir da avaliação de múltiplos datasets de STS, incluindo STS12-16, STSb e SICK-R. O SRoBERTa-NLI-large superou outros

modelos como SBERT-NLI-base e SBERT-NLI-large.

Nesse contexto, foi calculada a correlação de Spearman para os modelos pré-treinados após a aplicação do *fine-tuning* utilizando os corpora apresentados no estudo (Tabela 5.9). Essa métrica foi escolhida para complementar a análise realizada com a correlação de Pearson, como também para efeitos comparativos com os resultados obtidos no artigo do S-BERT. Os resultados confirmaram a mesma tendência observada anteriormente: o modelo pré-treinado com melhor desempenho foi o LaBSE, enquanto o modelo com pior desempenho foi o XLM-RoBERTa-base.Além disso, também se observou que os corpora SICK-BR e ASSIN 2 proporcionaram resultados superiores em comparação ao ASSIN, assim como mostrado com a correlação de Pearson.

Tabela 5.9 – Correlação de Spearman para diferentes modelos em ASSIN, ASSIN 2 e SICK-BR

Modelos	**ASSIN**	**ASSIN 2**	**SICK-BR**	**Média**
xlm-roberta-base	73.20	76.36	80.31	76.62
LaBSE	79.54	79.74	84.13	81.13
MiniLM-L12-v2	77.09	80.75	84.39	80.74
bert-base-portuguese	77.12	78.36	82.22	79.23

Os resultados exibidos no artigo (Reimers; Gurevych, b) se deram com diferentes corpora dos utilizados no presente estudo, o que não permite uma comparação totalmente fidedigna. O modelo que apresentou a melhor média em (Reimers; Gurevych, b) foi o modelo SRoBERTa-NLI-large (76,68 de correlação de Spearman), enquanto os modelos LaBSE (81.13),MiniLM-L12-v2 (80.74) e bert-base-portuguese (79,23) no presente estudo e com diferentes corpora superaram ligeiramente o desempenho do SRoBERTa-NLI-large, utilizando os corpora já mencionados. Já o bert-base-portuguese, com uma média de 76,56, apresentou um desempenho levemente inferior, enquanto o XLM-RoBERTa-base registrou o menor desempenho (76.62), destacando suas dificuldades em capturar nuances semânticas do português, mesmo após o fine-tuning.

O desempenho do bert-base-portuguese é especialmente relevante, pois trata-se do único modelo monolíngue avaliado, projetado exclusivamente para o português. Apesar de sua arquitetura mais simples e de ser treinado com um volume de dados menor em comparação com modelos multilinguagem, como o LaBSE, o bert-base-portuguese demonstrou um desempenho competitivo, sendo levemente mais eficaz que o estado da arte representado pelo SRoBERTa-NLI-large.

Esses resultados indicam uma oportunidade significativa para futuras pesquisas explorarem não apenas o desenvolvimento de modelos monolíngues mais robustos para o português, mas também comparações mais aprofundadas entre abordagens monolíngues e multilinguagem. Ao mesmo tempo, reforçam a necessidade de métricas diversificadas e

corpora especializados, como os utilizados neste estudo, para avaliar o impacto de diferentes arquiteturas em tarefas desafiadoras como STS.

6 Conclusão e sugestões

De acordo com a proposta inicial de comparação entre Modelos Multilíngues e Monolíngues para a tarefa de STS em português, foi observado que, apesar da expectativa inicial de que uma abordagem monolíngue proporcionaria um melhor desempenho, os modelos multilíngues apresentaram resultados levemente superiores. Esse desempenho pode ser atribuído a diversos fatores, como a maior robustez adquirida durante o pré-treinamento em múltiplos idiomas, a capacidade de generalização aprimorada e a exposição a um volume maior de dados textuais variados.

O presente estudo atingiu seu objetivo ao explorar essa comparação, destacando a eficiência dos modelos multilíngues na tarefa de STS em português, bem como a importância do *fine-tuning* em tarefas específicas e da escolha do modelo em aplicações que envolvem análise semântica textual.

O estudo de Processamento de Linguagem Natural tem se mostrado de extrema importância no contexto tecnológico atual, considerando que muitas ferramentas utilizadas no cotidiano dependem desses avanços para viabilizar uma interface de comunicação eficiente entre humanos e computadores. Tecnologias como assistentes virtuais, mecanismos de busca e sistemas de recomendação são exemplos diretos de aplicações que se beneficiam do PLN. Nesse cenário, o presente estudo buscou contribuir especificamente para a área de Similaridade Semântica Textual em português, uma vez que a maior parte das pesquisas relacionadas ao tema concentra-se em inglês ou outras línguas amplamente utilizadas, deixando o português relativamente menos explorado.

Embora o estudo tenha utilizado o modelo S-BERT, existem outros modelos de relevância que podem ser considerados em estudos futuros, como o GPT (Mao *et al.*, 2023). Esses modelos apresentam arquiteturas distintas e estratégias de pré-treinamento diferentes que podem oferecer vantagens ou desvantagens em cenários de STS. Uma proposta interessante seria realizar estudos comparativos detalhados entre o S-BERT e esses modelos no contexto da língua portuguesa. Esse tipo de análise poderia ajudar a identificar qual abordagem é mais eficiente em capturar nuances semânticas específicas do idioma, bem como em tarefas relacionadas, como agrupamento de sentenças, detecção de paráfrases e classificação de textos.

No presente estudo, não foi considerada a distribuição das classes de similaridade, como paráfrase, implicação e não-paráfrase, frequentemente utilizadas em tarefas de classificação semântica textual. A ausência dessa análise limita a compreensão do desempenho dos modelos em diferentes níveis de similaridade, uma vez que tais categorias permitem uma avaliação mais granular da capacidade dos sistemas em capturar relações semânticas específicas entre pares de sentenças. Recomenda-se, portanto, que trabalhos futuros incor-

porem essa dimensão classificatória, investigando como os modelos se comportam diante de cada classe e avaliando possíveis desequilíbrios na distribuição dos dados, o que pode impactar diretamente na acurácia e na generalização dos resultados. Essa abordagem pode fornecer insights mais precisos sobre os pontos fortes e limitações dos modelos de STS na língua portuguesa.

Uma direção promissora para trabalhos futuros é a avaliação dos modelos de Similaridade Textual Semântica (STS) em benchmarks amplamente utilizados, como o SNLI (Stanford Natural Language Inference), que fornece um cenário consolidado para tarefas de inferência textual

A observação de que modelos multilinguísticos, como o LaBSE e o MiniLM-L12-v2, superaram levemente o BERT-base-portuguese, apesar de este último ser especializado na língua portuguesa, sugere que a exposição a um volume maior e mais diverso de dados durante o pré-treinamento pode contribuir significativamente para a capacidade de generalização dos modelos em tarefas de STS. No entanto, essa superioridade não descarta o potencial de um modelo monolíngue mais competitivo. Uma solução viável seria o treinamento de um modelo monolíngue do zero (from scratch) em português, utilizando um corpus extenso e de alta qualidade, como o GIGAVERBO, com seus 200 bilhões de tokens. Além disso, o desenvolvimento de um benchmark específico e mais abrangente para o português, que incorpore diferentes tarefas semânticas (como inferência textual, paráfrase e similaridade por graus), seria essencial para avaliar adequadamente esses modelos e fomentar o avanço da área na língua estudada.

Uma possibilidade promissora para lidar com as limitações dos modelos multilíngues na captura de nuances morfológicas e expressões idiomáticas do português é a incorporação de abordagens baseadas em gramática no processo de tokenização e representação textual. Inspirando-se na proposta de GrammarCoder (Liang *et al.*, 2025), que demonstrou ganhos significativos em tarefas de geração de código ao empregar representações baseadas em regras gramaticais, sugere-se explorar métodos semelhantes na modelagem de linguagem natural. Ao representar sentenças com base em estruturas gramaticais — como árvores sintáticas e regras morfossintáticas — seria possível capturar de forma mais explícita as dependências linguísticas e as variações semânticas complexas do português. Essa abordagem poderia contribuir para uma diferenciação mais eficaz de sentidos e melhorar o desempenho em tarefas de STS, especialmente em casos que envolvem flexão verbal, concordância e construções idiomáticas que muitas vezes são sub-representadas em modelos treinados apenas com tokenização estatística.

Adicionalmente, os corpora utilizados neste trabalho, como o ASSIN, ASSIN 2 e SICK-BR, foram anotados não apenas para a tarefa de STS, mas também para Inferência Textual. Isso abre possibilidades para estudos futuros focados nesta tarefa em português. Avaliar como os modelos se comportam na identificação de relações de inferência, como implicação,

contradição ou neutralidade entre pares de sentenças, pode contribuir para o avanço do PLN no idioma. Além disso, seria valioso explorar as interseções entre STS e RTE, comparando o desempenho dos modelos em aplicações onde essas tarefas se sobrepõem, como em sistemas de perguntas e respostas, detecção de inconsistências em documentos e identificação de relações de causa e efeito em textos.

Como trabalho futuro, propõe-se a utilização do corpus GIGAVERBO, um recurso linguístico de larga escala totalmente em português, com aproximadamente 200 bilhões de tokens e 780 GB de texto, como base para o treinamento e avaliação de modelos de Similaridade Textual Semântica (STS). A adoção desse corpus apresenta vantagens significativas, como a ampla representatividade de registros linguísticos, domínios temáticos e variações regionais da língua portuguesa, o que pode contribuir para a construção de modelos mais robustos, generalizáveis e sensíveis às nuances semânticas do idioma. Além disso, o volume massivo de dados disponíveis permite o desenvolvimento de arquiteturas baseadas em aprendizado profundo com maior capacidade de generalização, o que pode impulsionar avanços no desempenho de tarefas de STS em português, ainda pouco exploradas em comparação com a língua inglesa.

Por fim, estudos futuros poderiam também expandir o uso dos corpora para incluir tarefas mais práticas e aplicadas, como sistemas de recomendação baseados em similaridade textual, análise de sentimentos combinada com STS, e modelos híbridos que unam a capacidade de inferência e similaridade em um único pipeline. Essas abordagens não apenas ampliariam o impacto das pesquisas em STS e RTE no português, mas também poderiam servir como base para o desenvolvimento de ferramentas que beneficiem diretamente a comunidade lusófona.

Referências

ARTETXE, M.; SCHWENK, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 7, p. 597–610, 2019. Disponível em: https://doi.org/10.1162/tacl_a_00288. Citado na p. 32.

CHANG, T. A.; ARNETT, C.; TU, Z.; BERGEN, B. K. **When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages**. arXiv, , [*s.d.*]. Disponível em: http://arxiv.org/abs/2311.09205. Citado na p. 14.

Common Crawl. **Common Crawl: Open Repository of Web Data**. 2025. Disponível em: https://commoncrawl.org. Citado na p. 30.

CONNEAU, A.; KHANDELWAL, K.; GOYAL, N.; CHAUDHARY, V.; WENZEK, G.; GUZMáN, F.; GRAVE, E.; OTT, M.; ZETTLEMOYER, L.; STOYANOV, V. **Unsupervised Cross-lingual Representation Learning at Scale**. arXiv, , [*s.d.*]. Disponível em: http://arxiv.org/abs/1911.02116. Citado nas pp. 14, 28 e 32.

COSTA, W. M. **Similaridade Semântica entre Acórdãos para Apoio na Formulação de Jurisprudência do TCU**. Dissertação (Dissertação de Mestrado Profissional) — Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Ciência da Computação, Brasília, Brasil, 2023. Citado na p. 13.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. [*s.d.*]. Citado nas pp. 13, 14 e 16.

FENG, F.; YANG, Y.; CER, D.; ARIVAZHAGAN, N.; WANG, W. **Language-agnostic BERT Sentence Embedding**. arXiv, , [*s.d.*]. Disponível em: http://arxiv.org/abs/2007.01852. Citado nas pp. 29 e 32.

FONSECA, E. R.; SANTOS, L. B. dos; CRISCUOLO, M.; ALUíSIO, S. M. Assin: Avaliação de similaridade semântica e inferência textual. *In*: UNIVERSIDADE DE SãO PAULO. **PROPOR 2016 - International Conference on the Computational Processing of Portuguese**. 2016. Disponível em: https://propor2016.di.fc.ul.pt. Citado nas pp. 14 e 26.

FONSECA, E. R.; SANTOS, L. B. dos; CRISCUOLO, M.; ALUíSIO, S. M. Visão geral da avaliação de similaridade semântica e inferência textual. **Linguamática**, v. 8, n. 2, p. 3–13, Dezembro 2016. ISSN 1647-0818. Proposta recebida em Setembro 2016 e aceite para publicação em Novembro 2016. Disponível em: https://news.google.com/. Citado nas pp. 13 e 17.

Google. **Google Colaboratory**. 2025. Disponível em: https://colab.research.google.com. Citado na p. 33.

Google News. **Google News**. 2025. Disponível em: https://news.google.com. Citado na p. 26.

HUANG, X.; PENG, H.; ZOU, D.; LIU, Z.; LI, J.; LIU, K.; WU, J.; SU, J.; YU, P. S. Cosent: Consistent sentence embedding via similarity ranking. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 32, p. 2800–2813, 2024. Citado na p. 33.

LIANG, Q.; ZHANG, Z.; SUN, Z.; LIN, Z.; LUO, Q.; XIAO, Y.; CHEN, Y.; ZHANG, Y.; ZHANG, H.; CHEN, B.; XIONG, Y. Grammar-based code representation: Is it a worthy pursuit for llms? **arXiv preprint arXiv:2503.05507**, 2025. Disponível em: https://arxiv.org/abs/2503.05507. Citado na p. 45.

LIDDY, E. D. Natural language processing. *In*: **Encyclopedia of Library and Information Science, 2nd Ed.** Marcel Decker, Inc., 2001. Disponível em: https://surface.syr.edu/istpub/63. Citado na p. 13.

MAO, R.; CHEN, G.; ZHANG, X.; GUERIN, F.; CAMBRIA, E. Gpteval: A survey on assessments of chatgpt and gpt-4. **arXiv preprint arXiv:2308.12488**, 2023. Citado nas pp. 16 e 44.

MARELLI, M.; BENTIVOGLI, L.; BARONI, M.; BERNARDI, R.; MENINI, S.; ZAMPARELLI, R. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *In*: **Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)**. Dublin, Ireland: Association for Computational Linguistics, 2014. p. 1–8. Disponível em: https://aclanthology.org/S14-2001. Citado na p. 28.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: An introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, 2011. Disponível em: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000464. Citado na p. 16.

QI, Y.; SACHAN, D.; FELIX, M.; PADMANABHAN, S.; NEUBIG, G. When and why are pre-trained word embeddings useful for neural machine translation? *In*: WALKER, M.; JI, H.; STENT, A. (Ed.). **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 529–535. Disponível em: https://aclanthology.org/N18-2084/. Citado na p. 32.

REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The ASSIN 2 shared task: A quick overview. *In*: QUARESMA, P.; VIEIRA, R.; ALUíSIO, S.; MONIZ, H.; BATISTA, F.; GONçALVES, T. (Ed.). **Computational Processing of the Portuguese Language**. Springer International Publishing, , [*s.d.*]. v. 12037, p. 406–412. ISBN 978-3-030-41504-4

978-3-030-41505-1. Series Title: Lecture Notes in Computer Science. Disponível em: http://link.springer.com/10.1007/978-3-030-41505-1_39. Citado nas pp. 14 e 27.

REAL, L.; RODRIGUES, A.; SILVA, A. V. E.; ALBIERO, B.; THALENBERG, B.; GUIDE, B.; SILVA, C.; LIMA, G. D. O.; CâMARA, I. C. S.; STANOJEVIć, M.; SOUZA, R.; PAIVA, V. D. SICK-BR: A portuguese corpus for inference. *In*: VILLAVICENCIO, A.; MOREIRA, V.; ABAD, A.; CASELI, H.; GAMALLO, P.; RAMISCH, C.; OLIVEIRA, H. G.; PAETZOLD, G. H. (Ed.). **Computational Processing of the Portuguese Language**. Springer International Publishing, , [*s.d.*]. v. 11122, p. 303–312. ISBN 978-3-319-99721-6 978-3-319-99722-3. Series Title: Lecture Notes in Computer Science. Disponível em: https://link.springer.com/10.1007/978-3-319-99722-3_31. Citado nas pp. 14 e 28.

REIMERS, N.; GUREVYCH, I. Making monolingual sentence embeddings multilingual using knowledge distillation. *In*: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Association for Computational Linguistics, , [*s.d.*]. p. 4512–4525. Disponível em: https://www.aclweb.org/anthology/2020.emnlp-main.365. Citado na p. 20.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *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, , [*s.d.*]. p. 3980–3990. Disponível em: https://www.aclweb.org/anthology/D19-1410. Citado nas pp. 15, 16, 19, 21, 25, 33, 41 e 42.

SCHWENK, H.; WENZEK, G.; EDUNOV, S.; GRAVE, E.; JOULIN, A. Ccmatrix: Mining billions of high-quality parallel sentences on the web. **arXiv preprint arXiv:1911.04944v2**, 2020. Disponível em: https://arxiv.org/abs/1911.04944. Citado na p. 32.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In*: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8. Citado na p. 32.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, ; POLOSUKHIN, I. Attention is all you need. [*s.d.*]. Citado nas pp. 9, 13, 16, 18 e 19.

WANG, W.; WEI, F.; DONG, L.; BAO, H.; YANG, N.; ZHOU, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. **arXiv preprint arXiv:2002.10957**, 2020. Disponível em: https://arxiv.org/abs/2002.10957. Citado na p. 32.

WENZEK, G.; LACHAUX, M.-A.; CONNEAU, A.; CHAUDHARY, V.; GUZMáN, F.; JOULIN, A.; GRAVE, E. **CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data**. arXiv, , [*s.d.*]. Disponível em: http://arxiv.org/abs/1911.00359. Citado nas pp. 29 e 32.

Wikimedia Foundation. **Wikipédia, a enciclopédia livre**. 2025. Disponível em: https://www.wikipedia.org. Citado na p. 30.

Wikipédia, a enciclopédia livre. **Coeficiente de correlação de Pearson — Wikipédia, a enciclopédia livre**. 2025. [Online; acesso em 23-jan-2025]. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson. Citado na p. 23.

Wikipédia, a enciclopédia livre. **Coeficiente de correlação de postos de Spearman — Wikipédia, a enciclopédia livre**. 2025. [Online; acesso em 23-jan-2025]. Disponível em: https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_postos_de_Spearman. Citado na p. 24.


UnB