UNIVERSIDADE FEDERAL FLUMINENSE

LUIZA CUNHA DE MENEZES

TÉCNICAS DE AUMENTO DE DADOS PARA APOIAR SISTEMAS DE AVALIAÇÃO AUTOMÁTICA DE LEITURABILIDADE

NITERÓI

2023

LUIZA CUNHA DE MENEZES

TÉCNICAS DE AUMENTO DE DADOS PARA APOIAR SISTEMAS DE AVALIAÇÃO AUTOMÁTICA DE LEITURABILIDADE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientadora:
ALINE MARINS PAES CARVALHO

Coorientadora:
MARIA JOSÉ BOCORNY FINATTO

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

C972t Cunha De Menezes, Luiza
TÉCNICAS DE AUMENTO DE DADOS PARA APOIAR SISTEMAS DE AVALIAÇÃO AUTOMÁTICA DE LEITURABILIDADE / Luiza Cunha De Menezes. - 2023.
107 f.: il.

Orientador: Aline Marins Paes Carvalho.
Coorientador: Maria José Bocorny Finatto.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2023.

1. Processamento de Linguagem Natural. 2. Aumento de Dados. 3. Substituição por Sinônimo. 4. Avaliação Automática de Leiturabilidade. 5. Produção intelectual. I. Marins Paes Carvalho, Aline, orientadora. II. Bocorny Finatto, Maria José, coorientadora. III. Universidade Federal Fluminense. Instituto de Computação.IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

LUIZA CUNHA DE MENEZES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em Março de 2023.

BANCA EXAMINADORA

Profa. Aline Marins Paes Carvalho - Orientadora, UFF

Profa. Maria José Bocorny Finatto - Coorientadora, UFRGS

Profa. Isabel Cristina Mello Rosseti, UFF

Profa. Maria Cláudia de Freitas, PUC-Rio

Niterói
2023

Gostaria de expressar minha gratidão a todos que contribuíram para a realização deste trabalho. Especialmente a minhas orientadoras, Aline e Maria José, pela orientação dedicada e pelos valiosos conselhos e sugestões que ajudaram a seguir com esta pesquisa. Também agradeço a minha família e amigos pelo apoio e incentivo incondicionais ao longo do meu percurso acadêmico. A todos, meu sincero agradecimento.

Resumo

O termo leiturabilidade traduz um conceito associado a uma condição-resultado do processo de simplificação textual (ST). Seu objetivo é transformar um texto de partida julgado complexo em algo de mais fácil compreensão para um dado tipo de leitor-principal. Estudos relacionados a como medir a leiturabilidade remontam o século passado. No entanto, não há um consenso entre estudiosos da área sobre as métricas existentes. Mais recentemente, tem-se avaliado o uso de ferramentas advindas do processamento de linguagem natural (PLN) para apoiar esta tarefa. Tais ferramentas dependem de um grande número de pontos no conjunto de dados de treino, o que implica em uma das maiores barreiras ao PLN avançado, que é o gargalo de aquisição de conhecimento linguístico. Assim, o objetivo deste trabalho é analisar o impacto de dois métodos agnósticos de aumento de dados (AD) para mitigar tal gargalo no que tange a avaliação automática por leiturabilidade (AAL) no português brasileiro (PB): a substituição por sinônimos (SS) e a retrotradução (RT). Para avaliação do impacto destes métodos, foram desenvolvidos 75 modelos para AAL, considerando diferentes seleções de entrada dos *corpora*, modelos classificatórios e seleções das representações de atributos de entrada. Em termos de *corpora*, foi estabelecido um *corpus* principal pareado e classificado, desenvolvido pela equipe de linguistas da UFRGS. E, em relação aos atributos, foram consideradas combinações de incorporações de palavras contextualizadas e/ou estáticas, e métricas de análise linguística e psicolinguística. Destaca-se que o melhor resultado obtido para o *corpus* principal sem aumento foi de 94,0% considerando um regressor logístico (RL) com incorporação de palavras contextualizadas (IPC). Este resultado foi melhorado para 95,2% ao combinar métricas de análise e IPC com aumento por RT da classe simples e SS de ambas as classes. Quando comparados a outros trabalhos no PB, a metodologia proposta gerou um aumento na acurácia do classificador em um domínio linguístico distinto ao de treino. Os resultados obtidos indicaram que os modelos treinados com uso de dados aumentados obtiveram uma capacidade igual ou superior àqueles treinados sem aumento e, ao mesmo tempo, apresentaram maior generalização quando introduzidos a outros domínios linguísticos.

Palavras-chave: Processamento de Linguagem Natural; Aumento de Dados; Substituição por Sinônimo; Retrotradução; Avaliação Automática de Leiturabilidade;

Abstract

The term readability translates to a concept associated with a condition-result of the textual simplification process. Its goal is to transform a starting text judged complex into something easier to understand for a given type of reader; studies about measuring readability date back to the last century. However, there is no consensus among specialists in the field about the existing metrics. More recently, tools derived from natural language processing (NLP) are supporting this task. Such tools depend on a high number of samples in the training dataset, which implies one of the barriers to advanced NLP, which is the bottleneck of acquiring linguistic knowledge. Thus, the objective of this work is to analyze the impact of two agnostic data augmentation methods to mitigate such bottlenecks in Brazilian Portuguese (BP) regarding automatic readability assessment (ARA): synonym substitution (SS) and back-translation (BT). To evaluate the impact of these methods, we developed seventy-five models for ARA, considering different selections of input *corpora*, classification models, and attribute representations. In terms of *corpora*, there is a main paired and annotated *corpus*, developed by the team of linguists at UFRGS; regarding attributes, the methodology considered combinations of contextualized and (/or) static word embeddings, and metrics of linguistic and psycholinguistic analysis. It is worth noting that the best result obtained for the main *corpus* without augmentation was 94.0% using logistic regression (LR) with contextualized word embeddings (CWE). This result improved to 95.2% by combining analysis metrics and CWE with augmentation by BT for the simple and SS for both simple and complex classes. Compared to other works in BP, the proposed methodology generated an increase in the classifier's accuracy in a different linguistic domain than the training one. The results obtained indicated that the models trained with augmented data obtained equal or superior capacity than those trained without augmentation and, at the same time, presented greater generalization when introduced to other linguistic domains.

Keywords: Natural Language Processing; Synonym Replacement; Back-translation; Data Augmentation; Automatic Readability Assessment;

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AAL Avaliação Automática de Leiturabilidade

AD Aumento de Dados

AM Aprendizado de Máquina

API *Application Programming Interface*

ARA *Automatic Readability Assessment*

BERT *Bidirectional Encoder Representations from Transformers*

BT *Backtranslation*

ET Elaboração Textual

EUA Estados Unidos da América

GPT *Generative Pre-trained Transformer*

IA Inteligência Artificial

INAF Indicador de Analfabetismo Funcional

LOO *Leave-one-out*

LSA Análise Semântica Latente

LSTM *Long Short-Term Memory*

ML Modelo de Linguagem

NILC Núcleo Interinstitucional de Linguística Computacional

ONG Organização Não-Governamental

PB Português Brasileiro

PLN Processamento de Linguagem Natural

RL Regressor Logístico

RNN *Recurrent Neural Network*

RT Retrotradução

SL Simplificação Léxica

SS Substituição por Sinônimo

ST Simplificação Textual

SVM *Support Vector Machine*

TTR Razão Tipo-*Token*

UFRGS Universidade Federal do Rio Grande do Sul

USP Universidade de São Paulo

WaC *Web as Corpus*

Sumário

1 Introdução

A comunicação é um elemento essencial na interação social humana. Para que ocorra de forma fluida e eficiente, é importante que locutor e interlocutor compartilhem do mesmo código, convenções e valores associados à linguagem. No entanto, as variações, tanto do código em si, como de interpretações destes códigos e valores, podem impedir que a comunicação aconteça de modo efetivo. Um exemplo comum de variação de código é dado pelos diferentes idiomas existentes. Se o locutor apresenta uma mensagem em português a um interlocutor que compreende apenas inglês, a comunicação será inviabilizada. Uma tarefa que facilita a remoção desta barreira é a tradução interlinguística, que corresponde à transformação de um texto fonte em um texto em outra língua, mantendo a fidelidade dos valores do texto original.

Analogamente à tradução interlinguística, tem-se, no contexto de acessibilidade textual, a Simplificação Textual (ST). Esta pode ser pensada como uma espécie de tradução intralinguística (FINATTO; TCACENCO, 2021) pois envolve tratar de uma adaptação em que se modifica um texto dentro da própria língua. Segundo Finatto (2020), a tarefa de ST tem por objetivo transformar um texto, escrito em uma língua A, que se presume que seja complexo, em uma versão, na mesma língua A, de modo que seja de mais fácil compreensão para um dado leitor-alvo. Ou seja, este processo de alteração é capaz de mitigar um possível ruído de comunicação, extrapolando-se a transformação que haveria se fosse o caso de uma tradução operando com dois idiomas diferentes.

Em condições normais, a dificuldade em interpretar textos está relacionada ao momento de aprendizado do interlocutor (que pode estar aprendendo um novo idioma, uma nova área de conhecimento, ou ainda estar em idade escolar). No entanto, a partir do momento em que um indivíduo adulto que é capaz de reconhecer letras e números, não consegue compreender textos simples em sua língua nativa e fazer operações matemáticas elementares, ele se torna um analfabeto funcional.

Segundo o Balanço do Plano Nacional de Educação, disponibilizado pela ONG Cam-

panha Nacional pelo Direito à Educação (CNDE, 2021), o Indicador de Analfabetismo Funcional (INAF) da população brasileira entre 15 e 64 anos aumentou de 27% em 2015 para 29% em 2018. Isto significa que cerca de 3 a cada 10 brasileiros têm uma capacidade limitada para validar informações e interpretar textos escritos. Conforme apontado por Leal (2019), um adulto que apresenta dificuldades de leitura dificilmente dispõe do tempo necessário para se dedicar a sua própria educação e, normalmente, é altamente dependente de um investimento financeiro que nem sempre é compatível com a sua renda.

Desta forma, esta realidade acaba por marginalizar uma grande parcela da população brasileira no que diz respeito ao acesso à informação e a outros direitos do cidadão. Pode-se dizer neste contexto, que a ST é uma tarefa que potencializa o acesso à informação e que, portanto, é importante tanto no processo de aprendizado, por facilitar o controle do nível de complexidade dos textos consumidos pelo estudante à medida que seu grau de proficiência evolui, como também por apresentar papel social relevante, ao facilitar o acesso à informação por meio de conteúdos coerentes.

Assim, a ST trata de algo bem mais sutil visto que está relacionada ao reconhecimento ou antecipação, por parte de quem escreve, das diferentes capacidades de interpretação do interlocutor. Neste contexto, a ST pode ser pensada como uma ferramenta de apoio à criação de conteúdos com nível de leiturabilidade acessível à habilidade do leitor-alvo (LEAL, 2019). De acordo com PONOMARENKO (2018), leiturabilidade é uma condição de facilidade de leitura criada por escolhas de conteúdo, estilo, *design* e organização que se alinham ao conhecimento prévio, escolaridade, interesse e motivação do público leitor. Logo, podemos inferir que a ST se relaciona diretamente à tarefa de quantificação da leiturabilidade de um texto; ou ainda que, sem a habilidade de classificar a leiturabilidade de um texto, não é possível desenvolver uma tarefa de ST.

1.1 Problema de Pesquisa

Atribuir um número à característica subjetiva da leiturabilidade tem sido motivo de estudos ao longo das últimas décadas, em diferentes idiomas (KINCAID; YASUTAKE; GEISELHART, 1967; KLARE et al., 1963; CAYLOR et al., 1973; MARTINS et al., 1996). Entretanto, até o momento, não há consenso entre os estudiosos da área de Linguística Descritiva ou Computacional em termos de estabelecer uma categorização ou escala numérica ideal para estimar a leiturabilidade textual. Isto porque as métricas existentes se relacionam às propriedades estruturais dos textos, e tendem a não considerar elemen-

tos como a interação entre texto e leitor (SANTOS, 2010; SCARTON; ALUÍSIO, 2010; FINATTO, 2020).

Segundo Taylor (1953), fórmulas assumem uma alta correlação entre a facilidade de compreensão e a frequência de ocorrência de tipos selecionados de elementos de linguagem, como, por exemplo, palavras curtas ou comuns, frases curtas ou simples, presença de certas partes do discurso, voz ativa ou passiva e outros. Uma alternativa para mitigar o impacto da superficialidade de tais propriedades pode ser obtida por meio da adoção de métodos de Processamento de Linguagem Natural (PLN) (MARTIN, 2009; GOLDBERG, 2016; HEATON, 2018; EISENSTEIN, 2018; FREITAS, 2022), que consideram uma abordagem probabilística para modelar a linguagem e a semântica. Conforme apontado por Freitas (2022), nos últimos anos, a aplicação de técnicas de Aprendizado de Máquina (AM), ou, ainda mais especificamente, o aprendizado profundo - em inglês, *deep learning* (LECUN; BENGIO; HINTON, 2015), ao PLN tem trazido resultados notoriamente positivos. Assim, com os avanços em PLN, em termos de um enfoque computacional para as métricas supracitadas, a tarefa de avaliação estimativa de leiturabilidade recebe o nome de *Automatic Readability Assessment* (ARA) ou, em português, Avaliação Automática de Leiturabilidade (AAL).

Neste contexto, pesquisadores que atuam em Linguística Computacional investigam o uso de modelos de língua estatísticos para tentar capturar características linguísticas e desenvolver modelos de AM para classificação automatizada de textos quanto à sua leiturabilidade. Isso é feito considerando-se a identificação e quantificação de elementos da superfície dos textos escritos, diferentes cenários de comunicação e variados perfis de leitores. Observa-se, portanto, que a realização automática de tarefas relacionadas ao domínio da comunicação e usos da linguagem é intrinsecamente complexa, estendem-se a estes desafios questões relacionadas à disponibilidade, geralmente escassa, de conjuntos robustos de dados linguísticos, coerentes e classificados para o problema analisado. Isto porque, conforme apontado por Brill (2003), Şahin (2022) e Freitas (2022), uma das maiores barreiras ao PLN avançado é o gargalo de aquisição de conhecimento linguístico via *corpora* textuais, tendo em vista que os modelos atuais de PLN dependem de um número elevado de amostras no conjunto de dados de treinamento, pois, sem dados robustos, eles não são capazes de generalizar a tarefa além dos dados de treinamento. Uma necessidade é a anotação dos *corpora*, que precisam trazer uma camada de informação linguística sobre as palavras que os perfazem, como informações semânticas e sintáticas, por exemplo.

Especificamente no que tange o Português Brasileiro (PB), a ausência de um grande

conjunto de textos previamente pareado em termos de leiturabilidade [1] intensifica ainda mais os desafios supracitados. Embora seja possível argumentar que atualmente a *web* é considerada uma fonte extraordinária de textos dos mais variados domínios e idiomas em formato digital, o trabalho associado à classificação dos textos em termos de leiturabilidade ainda é uma tarefa analítica, complexa e alvo de discussão entre especialistas (FINATTO; PARAGUASSU, 2022), que deve considerar as especificidades e características próprias de cada idioma.

Com o intuito de melhorar o desempenho de sistemas automáticos de PLN frente à carência de grandes conjuntos de dados provenientes de usos de linguagem, pesquisadores têm introduzido métodos de Aumento de Dados (AD) no domínio de análise e processamento textual, com a finalidade de aumentar o tamanho de uma dada amostra a ser utilizada em um treinamento (FENG; GANGAL et al., 2021; BAYER; KAUFHOLD; REUTER, 2021). Destaca-se, no entanto, que há uma lacuna na literatura quando o assunto é a classificação textual para estimar a leiturabilidade e o uso de ferramentas que sejam capazes de mitigar o gargalo da falta de grandes quantidades de dados linguísticos. De modo que não foram localizados trabalhos que propusessem o uso de AD para *corpora* em PB; e além disso, não foram identificados trabalhos em outros idiomas que propusessem o uso de AD especificamente voltados para a temática da leiturabilidade.

Neste sentido, uma questão primordial surge: *como reduzir o gargalo de informação de textos classificados por leiturabilidade na língua portuguesa?* A resposta a esta pergunta pode fornecer uma fonte numerosa e altamente qualificada de *corpora* para inúmeras tarefas do domínio de acessibilidade textual. Na seção seguinte (1.2), uma alternativa é proposta e esta será explorada ao longo deste trabalho.

1.2 Formulação de hipótese e objetivos

Dado o contexto mencionado na seção 1.1, o trabalho aqui relatado busca responder à questão supracitada por meio do aumento artificial a partir de um conjunto inicial de textos em um *corpus* classificado em termos de graus de leiturabilidade e atendendo às particularidades do PB. Ou seja, a partir de um número reduzido de textos, criam-se novos textos sem a necessidade de coleta e classificação de outros textos.

Assim, a hipótese de pesquisa é que o uso de determinados métodos de AD pode

[1] Por pareado compreende-se que estarão presentes no *corpus* pares de texto: (1) um em linguagem simples, (2) e outro como sendo uma representação muito próxima de (1) em termos de conteúdo, fazendo uso de uma linguagem prolixa ou altamente especializada em um domínio linguístico.

prover textos sintéticos aderentes ao contexto e com volumetria suficiente para reduzir o impacto do gargalo de informação para o treinamento de modelos de AAL para o PB. Mais especificamente, são explorados os métodos de Substituição por Sinônimo (SS) e Retrotradução (RT) devido à necessidade de manutenção da classe original do texto. Tais métodos apresentam alta capacidade de parafraseamento e de preservação de conteúdo, e, ao mesmo tempo, por serem tarefas agnósticas, podem ser adaptadas facilmente para diferentes domínios com menor custo computacional.

Como forma de corroborar ou refutar a hipótese, a metodologia aplicada inclui uma análise quantitativa comparativa do efeito da inclusão de exemplos gerados artificialmente no treinamento do processo de classificação. Por isso, foi realizada a criação de um classificador textual binário automático em termos de leiturabilidade, *i.e.*, capaz de classificar um texto como simples ou não. No contexto de modelos de classificação, a escolha dos atributos de entrada é uma parte essencial do processo de AAL. Dentre os atributos disponíveis atualmente, foram exploradas combinações de diferentes possíveis representações numérico-textuais relacionados direta ou indiretamente ao domínio da leiturabilidade. Assim, foram consideradas métricas de análise linguística e psicolinguística disponíveis para o PB gerados pelo NILC-Metrix (LEAL; SCARTON et al., 2022) e representações por incorporação de palavras (ou no inglês, *word embeddings* (MIKOLOV et al., 2013)) tanto estáticas quanto contextualizadas.

Neste sentido, as questões de pesquisa que este trabalho busca responder são:

- Qual o efeito da inclusão de exemplos gerados artificialmente para o processo de classificação por AAL?
- Quais atributos de entrada de um modelo de AAL por AM produzem resultados mais aderentes com o problema da classificação por leiturabilidade?
- Considerando um domínio específico D_w com poucos exemplos de textos simplificados, e um classificador C_k para decidir se um texto é 'simples' ou não, construído a partir de dados de domínios genéricos $D_g = \{D_1, \cdots, D_n\}$ com uso de métodos de AD. C_k pode ser generalizável para o domínio específico $D_w \notin D_g$?

Em suma, este trabalho tem como objetivo geral identificar empiricamente determinados métodos de AD que se comprovem válidos no contexto de AAL. Para contribuir com o alcance do objetivo principal, definem-se como objetivos secundários deste trabalho:

1. desenvolvimento de métodos de AD aplicados especificamente para o PB a partir

dos métodos de SS e RT;

2. geração de um *corpus* pareado por meio do uso dos métodos de AD desenvolvidos, a partir de um conjunto reduzido e previamente classificado de textos;

3. criação de um classificador binário textual automático de leiturabilidade.

Destaca-se que os *corpora* utilizados, bem como todo desenvolvimento realizado tanto dos métodos de AD quanto dos modelos de classificação estão disponíveis para *download* no repositório MeLLL-UFF no Github [2].

A seguir, no Capítulo 2, conceitualizamos o termo de leiturabilidade, bem como a área de AD, e também demonstramos os conceitos relacionados às escolhas das representações de atributos textuais para conversão numérica utilizados no processo de análise da proposta de AD desenvolvida. Na sequência, no Capítulo 3, apresentamos trabalhos relacionados à tarefa de AAL; no Capítulo 4, delineamos a proposta de AD desenvolvida, considerando exemplos de implementação; e, no Capítulo 5, expomos a metodologia experimental com maior detalhamento das etapas embutidas. Por fim, elucidamos os resultados obtidos pelo experimento realizado no Capítulo 6, e apresentamos as conclusões com indicações de trabalhos futuros e considerações finais no Capítulo 7.

[2] https://github.com/MeLLL-UFF/text-simplification/tree/dissertacao_luiza-menezes/aumento_de_dados_classificacao

2 Fundamentação Teórica

Neste Capítulo são apresentados os conceitos fundamentais para embasamento da metodologia desenvolvida e da interpretação dos resultados obtidos. Para facilitar a leitura, foram consideradas três seções apresentadas a seguir: (1) Leiturabilidade, apresenta a relevância do tema por meio de uma linha do tempo com alguns dos principais trabalhos sobre o assunto; (2) AD, sintetiza e classifica diversos métodos de AD aplicados ao PLN para que o leitor possa observar o número de possibilidades disponíveis e entender, posteriormente as escolhas realizadas; e, (3) Representação de Atributos, descreve a evolução das principais estratégias utilizadas para representar textos de forma numérica, de modo que sirvam de entrada para os métodos de AM.

2.1 Leiturabilidade

Segundo Finatto e Paraguassu (2022), o termo leiturabilidade, adaptação para PB do termo em inglês *readability*, está associado a uma condição-resultado do processo de ST, cujo objetivo é tornar o texto de mais fácil compreensão e, portanto, mais acessível para o leitor-alvo. Neste contexto, entende-se por complexidade a condição ou estado de uma avaliação a respeito do potencial de leiturabilidade de um texto. Em uma visão simplista, pode-se dizer que, ao avaliar o nível de complexidade de um texto é possível considerar uma técnica de ST para garantir maior leiturabilidade do público-leitor para o qual o texto se destina.

Uma outra tarefa do domínio da acessibilidade textual que não deve ser confundida com a ST, é a Elaboração Textual (ET). Enquanto a primeira tem impacto direto na leiturabilidade do texto, a ET impacta na compreensibilidade. Ou seja, enquanto a ST busca adaptar a complexidade lexical ou sintática do texto para o leitor-alvo, as tarefas relacionadas à ET buscam inserir conteúdos explicativos ao longo do texto com o intuito de expandir o vocabulário do leitor (HARTMANN; ALUÍSIO, 2020).

Neste trabalho, a leiturabilidade será tratada como uma indicação do quão fácil ou

difícil um texto é. No entanto, atribuir um número, binário ou não, à característica subjetiva da leiturabilidade tem sido motivo de estudo ao longo das últimas décadas em diferentes idiomas. A fim de demonstrar a importância e evolução de alguns dos principais estudos de avaliação deste índice, com foco principalmente no PB, a Tabela 1 foi preparada. A Tabela 1 foi organizada em três colunas com o intuito de se aproximar de uma linha do tempo. A primeira coluna delimita um período ou uma data exata de uma ou mais publicações relevantes; a segunda coluna apresenta o título da(s) publicação(ões) de interesse, de modo que os títulos em idiomas diferentes do PB foram traduzidos livremente pela autora; e, a última coluna apresenta um breve contexto do período.

Tabela 1: Evolução da avaliação de leiturabilidade textual

Período	Referência(s)	Resumo
1902	Casos curiosos: Uma coleção de decisões americanas e inglesas selecionadas por leiturabilidade (MILBURN, 1902)	Salvo melhor juízo, o livro "Casos curiosos" foi a referência mais antiga a mencionar o termo leiturabilidade. Faz parte de uma coleção de livros sobre o sistema legislativo contendo casos de estudo nos Estados Unidos da América (EUA) e Reino Unido.
1921	O livro de palavras do professor (THORNDIKE, 1921)	Segundo Carrell (1987), trabalhos buscando quantificar a leiturabilidade começaram entre 1915 e 1920, mas foi Thorndike (1921) quem encorajou o uso de diretrizes para níveis de ensino, fornecendo cerca de 10.000 palavras impressas em textos de amostra na língua inglesa.
1923	Método para medir a carga de vocabulário de livros didáticos (LIVELY; PRESSEY, 1923)	Segundo (CARRELL, 1987), a partir das frequências de palavras apresentadas por Thorndike (1921), Lively e Pressey (1923) apresentaram a primeira tentativa de estabelecer uma fórmula para a leiturabilidade considerando diversos índices para pontuar a dificuldade de um texto.

Tabela 1: Evolução da avaliação de leiturabilidade textual (Continuação)

Período	Referência(s)	Resumo
1935	O que torna um livro legível (GRAY; LEARY, 1935)	Neste trabalho, os autores entrevistaram 288 adultos envolvidos com educação e definiram quatro fatores para a leiturabilidade na língua inglesa: conteúdo, estilo, formato e organização (CARRELL, 1987). No entanto, propuseram fórmulas para a leiturabilidade apenas relacionadas ao estilo, padrão que foi seguido nas publicações seguintes.
1936 - 1974	Um novo critério de leiturabilidade (FLESCH, 1948) | Uso do índice de leiturabilidade automatizado das ordens técnicas da Força Aérea (KINCAID; YASUTAKE; GEISELHART, 1967) | Leiturabilidade automatizada para uso com materiais técnicos (SMITH; KINCAID, 1970) | Medição de leiturabilidade (KLARE et al., 1963) | Metodologias para leiturabilidade de especialidades ocupacionais militares (CAYLOR et al., 1973)	Segundo Carrell (1987), neste período, houve a necessidade de utilização de fórmulas de leiturabilidade nos EUA não apenas por questões educacionais, mas também para produção de materiais técnico-científicos durante a segunda guerra mundial. Em 1974, já eram consideradas mais de 200 variáveis de linguagem/estilo diferentes e havia quase tantas fórmulas diferentes.

Tabela 1: Evolução da avaliação de leiturabilidade textual (Continuação)

Período	Referência(s)	Resumo
1953	"Cloze Procedure": uma nova ferramenta para medir a leiturabilidade (TAYLOR, 1953)	Em meio às críticas sobre o uso de fórmulas para leiturabilidade que se iniciaram na década de 50, este artigo pode ser considerado precursor do que atualmente chamamos de modelos de linguagem mascarados(DEVLIN et al., 2019). Isto porque o modelo proposto considera a probabilidade de padrões de escrita com base na expectativa gerada pela leitura. Em outras palavras, Taylor (1953) propõe que a classificação de um texto em simples ou complexo está atrelada à expectativa atendida do leitor em descobrir palavras omitidas aleatoriamente no texto. É uma quebra de paradigma interessante haja vista que os modelos anteriores se baseavam majoritariamente em fórmulas.
1996	Fórmulas de leiturabilidade aplicadas a livros didáticos em português brasileiro (MARTINS et al., 1996)	Surge a primeira adaptação para avaliação de leiturabilidade textual disponível para o PB, inspirado no índice *Flesch-Kincaid Grade Level*, que avalia a complexidade de textos em uma escala de quatro níveis. Scarton e Aluísio (2010) mencionam que a fórmula para PB foi adaptada de modo que cada nível varia entre duas séries da educação primária, uma da educação secundária, e outra do ensino superior.

Tabela 1: Evolução da avaliação de leiturabilidade textual (Continuação)

Período	Referência(s)	Resumo
2001	Modelo estatístico para leiturabilidade científica (SI; CALLAN, 2001)	A partir dos anos 2000, com os avanços em PLN, a tarefa de avaliação de leiturabilidade evoluiu para o que conhecemos por ARA, ou AAL, em português. Pesquisadores em linguística computacional começaram a investigar o uso de *parsers*[1] e modelos estatísticos de linguagem para capturar características linguísticas e desenvolver modelos de AM para classificação de textos por leiturabilidade. No trabalho de Si e Callan (2001), modelos unigrama[2] são utilizados para classificar a leiturabilidade de páginas da *web* em inglês.
2005	Avaliação de leiturabilidade usando *Support Vector Machine* (SVM) e modelos estatísticos de linguagem (SCHWARM; OSTENDORF, 2005)	Com o trabalho de Schwarm e Ostendorf (2005), recursos sintáticos, extraídos por *parser*[3] foram utilizados para avaliação da leiturabilidade na língua inglesa. Verificou-se que combinações de recursos lexicais e sintáticos produzia melhores resultados para a tarefa de avaliação de leiturabilidade.
2006	Coh-Metrix: Pontuações automatizadas de coesão e coerência para prever a leiturabilidade e facilitar compreensão	Foi apresentada a ferramenta Coh-Metrix, capaz de calcular automaticamente métricas relevantes para a compreensão de textos em inglês com base em recursos como repetição de itens

[1]Um *parser* sintático consiste em uma análise automática da estrutura sintática de uma linguagem, tanto em termos das dependências gramaticais quanto das partes constituintes.

[2]Tais modelos assumem que a probabilidade de gerar uma palavra independe de seu contexto (SI; CALLAN, 2001).

[3]O *parser* sintático utilizado em Schwarm e Ostendorf (2005) foi inspirado no conceito de máxima entropia do trabalho de Charniak (2000).

Tabela 1: Evolução da avaliação de leiturabilidade textual (Continuação)

Período	Referência(s)	Resumo
	(MCNAMARA DS.; GRAESSER, 2002)	lexicais em frases e análise semântica latente (MCNAMARA DS.; GRAESSER, 2002).
2008	Rumo aos Sistemas Automáticos de Simplificação Textual do PB (ALUÍSIO et al., 2008)	Em termos de acessibilidade textual e de promoção da inclusão digital no PB, destaca-se o projeto PorSimples (ALUÍSIO et al., 2008). Este projeto tem por objetivo facilitar a compreensão de informações para crianças e adultos em processo de alfabetização ou pessoas com algum tipo de deficiência de leitura. Ao longo dos últimos anos, como parte do projeto, foram criadas duas ferramentas computacionais: o Facilita (WATANABE et al., 2009), que é um *plug-in* de navegador para simplificar sites automaticamente, e o Simplifica (SCARTON; OLIVEIRA et al., 2010), um editor destinado a produtores de conteúdo que desejam criar textos simplificados adequados ao mesmo público.
2009	Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português (ALMEIDA; ALUÍSIO, 2009)	Surge a primeira adaptação do Coh-Metrix para o PB, o Coh-Metrix-PORT (ALMEIDA; ALUÍSIO, 2009), que contava com 41 métricas. O trabalho descrito neste artigo faz parte do projeto PorSimples e nele foram mapeados diversos recursos de PLN disponíveis para o PB.

Tabela 1: Evolução da avaliação de leiturabilidade textual (Continuação)

Período	Referência(s)	Resumo
		Destaca-se o *parser* PALAVRAS (BICK, 2000), que continua sendo uma referência para o PB até os dias atuais e que está sendo constantemente melhorado.
2019	Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural (LEAL; MAGALHAES et al., 2019)	Voltada para a classificação de sentenças no PB foi implementada a interface de código aberto Simpligo (LEAL, 2020), na qual é possível conferir os valores preditos para cada sentença, numa escala de complexidade entre 0 e 100.
2021	NILC-Metrix: avaliando a complexidade de escrita e língua falada em PB (LEAL; SANCHES DURAN et al., 2021)	O Coh-Metrix-Port evoluiu ao longo dos últimos anos pelo Núcleo Interinstitucional de Linguística Computacional (NILC) e hoje, o NILC-Metrix (LEAL; SANCHES DURAN et al., 2021; LEAL; SCARTON et al., 2022) é considerado um dos recursos mais completos da área de PLN adaptados para o PB e disponíveis atualmente. A ferramenta é capaz de calcular 200 métricas de análise linguística e psicolinguística para o PB, capturando características de coesão e traços marcadores da dificuldade de um texto em vários níveis (léxico, sintático, discursivo e conceitual) a partir da integração de vários recursos e ferramentas, envolvendo léxicos, *taggers*, *parsers*, lista de marcadores discursivos, entre outros.

Ressalta-se que, principalmente, em relação aos idiomas com mais recursos (como a língua inglesa), a partir dos anos 2000, com o crescimento da disponibilidade de dados e o aumento da capacidade computacional, os algoritmos de AM têm se tornado cada vez mais precisos e eficazes. Isto significa que diversos trabalhos foram publicados ao longo dos últimos anos com técnicas de AM para AAL, e muitos desses estudos têm mostrado resultados promissores. Estes estudos permeiam tanto a construção de *corpora* classificados, como é o caso de Feng, Elhadad e Huenerfauth (2009), Vajjala e Lučić (2018) e Crossley et al. (2022), quanto o desenvolvimento de redes neurais profundas e o uso de modelos pré-treinados de linguagem.

Recentemente, modelos de linguagem pré-treinados *Transformer* (VASWANI et al., 2017; PETERS et al., 2018; RADFORD et al., 2018) foram relatados como superando notoriamente atributos manualmente elaborados (IMPERIAL, 2021; LEE; JANG; LEE, 2021; MARTINC; POLLAK; ROBNIK-ŠIKONJA, 2021). Mais informações sobre a tarefa de AAL podem ser encontradas no Capítulo 3.

2.2 Aumento de Dados (AD)

A abordagem de AD, ou *data augmentation*, em inglês, tem por objetivo o aumento da variedade de um conjunto inicial de dados sem a necessidade de nova coleta e categorização (FENG; GANGAL et al., 2021). Conforme esclarecido por Feng, Gangal et al. (2021), o AD deve fornecer uma alternativa para obtenção de mais dados, de modo que, um método ideal deve ser fácil de implementar e capaz de melhorar o desempenho do modelo. Destaca-se, no entanto, que Feng, Gangal et al. (2021) mencionam o fato de que há uma lacuna teórica que fundamente os estudos em AD, segundo os autores, a maioria dos estudos pode mostrar empiricamente que uma técnica de AD funciona, mas é um desafio medir a qualidade de uma técnica sem recorrer a um experimento em grande escala.

Os métodos de AD tiveram sua origem no campo da visão computacional (BAYER; KAUFHOLD; REUTER, 2021). Com imagens, transformações simples como cortar, girar e variar suas cores são úteis, pois permitem criar variações e particularidades que facilitam a generalização dos modelos de AM. Isto significa que as operações aplicadas às imagens normalmente não alteram a natureza daquilo que foi capturado, apenas as reapresentam em formatos diferentes. No domínio de PLN, a geração de exemplos aumentados de textos escritos, capazes de capturar as invariâncias e alternâncias linguísticas desejadas, é uma tarefa bem menos óbvia (FENG; GANGAL et al., 2021; FERREIRA; COSTA, 2020). Ao

ponderarmos alterações análogas às supracitadas para imagens em um texto, as estratégias intuitivas seriam a remoção de trechos em sentenças, reordenação ou inserção aleatória de palavras. Entretanto, tais mudanças implicam na possível violação de uma série de regras léxicas e sintáticas, além do distanciamento do conteúdo inicial em termos do significado de palavras e do significado do todo do texto. No entanto, conforme pontuado por Feng, Gangal et al. (2021), é possível se inspirar nos métodos de AD em imagens para textos. Por exemplo, aplicar uma escala de cinza em uma imagem colorida pode ser entendida como uma atenuação de aspectos linguísticos, como uma mudança do grau superlativo dos adjetivos contidos no texto.

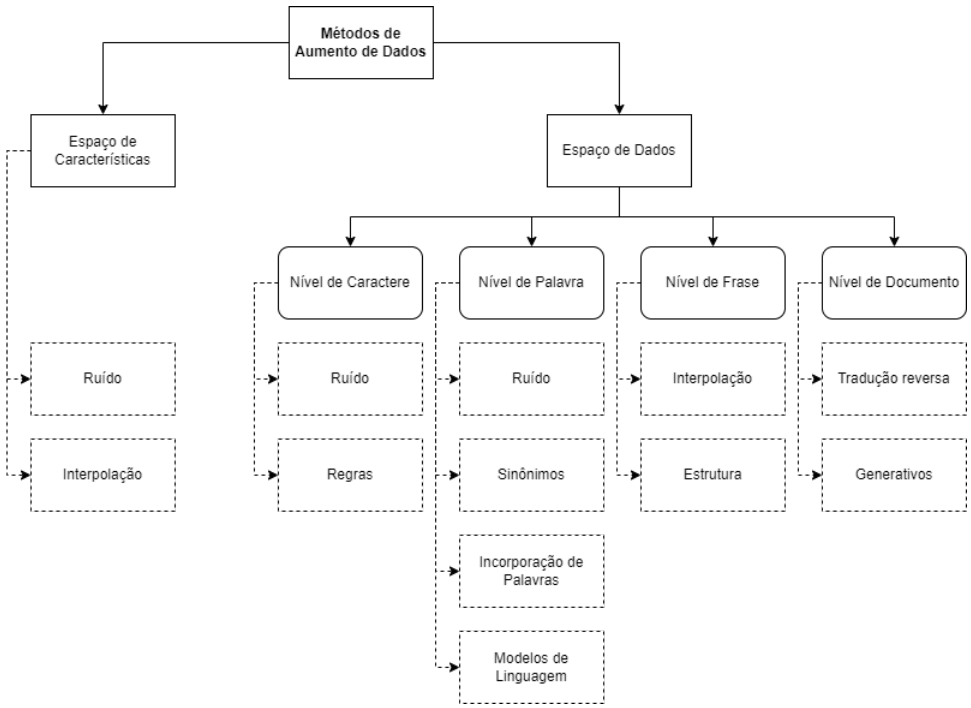


Figura 1: Taxonomia para diferentes métodos de AD. Adaptado de Bayer, Kaufhold e Reuter (2021)

Especificamente sobre o domínio de PLN, Bayer, Kaufhold e Reuter (2021) apontam que pesquisas conduzidas para o AD em PLN são recentes e eram escassas até 2019. No último ano, duas revisões da literatura no âmbito de AD para PLN foram disponibilizadas: (1) Feng, Gangal et al. (2021) e (2) Bayer, Kaufhold e Reuter (2021). O trabalho de Feng, Gangal et al. (2021) apresenta uma perspectiva mais geral sobre o AD em PLN com uma visão orientada por tarefas, enquanto o trabalho de Bayer, Kaufhold e Reuter (2021) se concentra na contextualização dos métodos de AD, comparando a origem e desempenho entre modelos, em uma visão orientada para o método. Em Bayer, Kaufhold e Reuter (2021), é proposta uma adaptação para PLN da taxonomia apresentada por Shorten e Khoshgoftaar (2019), vide Figura 1, em que os métodos de AD podem ser divididos

em: (a) espaço de dados, referente às transformações em dados brutos e (b) espaço de características, referente às transformações em representações de atributos da entrada.

Na Tabela 2 é possível verificar definições sintetizadas, fundamentadas no trabalho de Bayer, Kaufhold e Reuter (2021), para os métodos de AD representados na Figura 1.

Tabela 2: Principais métodos de AD aplicáveis para PLN

Espaço	**Nível**	**Método**	**Descrição**
Dados	Caractere	Indução por ruído	A ideia básica é adicionar ruído artificial aos dados de treinamento, a partir de operações como troca aleatória de letras únicas, ou de um sequência de caracteres dentro de uma palavra, bem como a exclusão ou inserção ou substituição de letras ou sequências de caracteres por outras.
		Substituição baseada em regras	As transformações válidas são, entre outras, a inserção de erros ortográficos, alterações de dados, nomes de entidades e abreviaturas.
	Palavra	Substituição por Sinônimo (SS)	É um método bastante popular de AD que parafraseia instâncias de texto substituindo certas palavras por sinônimos.
		Indução por ruído	Também é possível de ser aplicada ao nível de palavra. No ruído unigrama, as palavras nos dados de entrada podem ser excluídas, reduzidas ou substituídas por outra(s) palavra(s) dada uma certa probabilidade.
		Substituições por incorporações de palavras ou *word embeddings*	Este método se aproxima da SS, haja vista que busca substituir palavras por outras; neste caso, as palavras são traduzidas para um espaço de representação latente em que palavras com contextos semelhantes

Tabela 2: Principais métodos de AD aplicáveis para PLN (Continuação)

Espaço	**Nível**	**Método**	**Descrição**
			se aproximam em termos de posicionamento vetorial. Destaca-se que esta substituição não necessariamente garante a preservação da classe original das instâncias. Por exemplo, "o filme foi fantástico" e "o filme foi horrível" podem ser consideradas transformações válidas por estarem próximas no espaço de contexto.
		Substituições por Modelo de Linguagem (ML)	Diferente da substituição por incorporação de palavras, um ML permite uma substituição mais contextualizada porque considera uma previsão de palavra dada uma sequência. Se a sequência de palavras utilizada no modelo for anterior à palavra que se deseja prever/substituir, trata-se de um ML clássico; se a sequência for bidirecional, *i.e.*, circundante, trata-se de um ML mascarado. Ressalta-se que, da mesma forma que a substituição por incorporação de palavras, este método pode alterar a classe original do dado aumentado. Isto porque, não necessariamente as palavras substituídas terão um significado similar às do texto original, uma vez que as substituições por ML consideram a distribuição estatística de sequências de palavras.

Tabela 2: Principais métodos de AD aplicáveis para PLN (Continuação)

Espaço	**Nível**	**Método**	**Descrição**
	Frase	Interpolação	Ainda que a interpolação seja pertencente ao espaço de características, o estudo de Shi, Livescu e Gimpel (2021) é considerado por Bayer, Kaufhold e Reuter (2021) uma aproximação da interpolação no espaço de dados. Shi, Livescu e Gimpel (2021) interpolam instâncias do espaço de dados por meio de substituições de subsequências de palavras por outras com mesma classe morfossintática e rótulo. Para escolha da substituição, uma combinação de uma série de regras podem ou não ser aplicadas.
		Transformação baseada em estrutura	Esta abordagem é limitada a certos idiomas ou tarefas pois dependem de pré-processamento de estrutura de conteúdo, por meio de técnicas como marcadores de discurso[4], *parsers*[5], e outros. Um exemplo de transformação relacionado a este método é a mudança da voz ativa para passiva de uma sentença e vice-versa, o que pode ser considerada uma aproximação da rotação de uma imagem no domínio visual.

[4]Em inglês *Part-of-Speech (POS) tagging*, é um componente de PLN que provê a categoria lexical das palavras em uma sentença.

[5]Para fins de esclarecimento, um *parser* sintático considera a sintaxe das sentenças, *i.e.*, a função que as palavras desempenham dentro da oração (e.g. sujeito, objeto direto, complemento nominal, e outros). Um *POS tagger* analisa a classe gramatical que as palavras da sentença se referem (e.g., substantivo, adjetivo, verbo, advérbio e outras).

Tabela 2: Principais métodos de AD aplicáveis para PLN (Continuação)

Espaço	**Nível**	**Método**	**Descrição**
	Documento	Retrotradução (RT) ou Tradução reversa	É uma abordagem para parafrasear texto com a ajuda de modelos de tradução, de modo que o documento original é traduzido para outro idioma (tradução direta) e depois traduzido de volta para o idioma de origem (tradução reversa). Este tipo de método funciona porque tradutores automáticos avaliam diversas combinações na escolha de termos ou de estrutura da frases que variam conforme a complexidade da linguagem.
		Métodos generativos	Bayer, Kaufhold e Reuter (2021) apresentam uma série de estudos cujo propósito é gerar textos artificiais totalmente novos. Tais estudos consideram o uso de redes neurais com diferentes arquiteturas, como *Recurrent Neural Network* (RNN), adversárias generativas (GAN), *autoencoders* variacionais (VAE), e outras. Os métodos generativos por si só podem ser considerados uma subárea de PLN, no entanto, podem ser utilizados com a finalidade de aumentar o número de dados de um conjunto. É importante atentar que são métodos altamente custosos computacionalmente e de difícil preservação de rótulos.

Tabela 2: Principais métodos de AD aplicáveis para PLN (Continuação)

Espaço	Nível	Método	Descrição
Características	N/A[6]	Indução por ruído	Da mesma maneira que no espaço de dados, ruídos podem ser inseridos no espaço de características. É possível realizar operações aleatórias de multiplicação, adição, substração, divisão no espaço de representação latente.
	N/A	Interpolação	Os métodos de interpolação neste espaço consistem na combinação de estados de sentenças independentes para criação de novos estados que contêm significado das sentenças originais. Ressalta-se que, conforme apontado por Bayer, Kaufhold e Reuter (2021), a transformação de retorno para o espaço de dados pós interpolação não é trivial.

É importante salientar que o uso de métodos de AD para PLN ainda apresenta resultados limitados em termos de ganhos de desempenho, e como consequência, esses métodos automáticos têm sido pouco explorados pela comunidade. Isso vem ocasionando, por exemplo, uma carência de compreensão entre AD e implicações no modelo de aprendizado (FERREIRA; COSTA, 2020). Enquanto isso, o estudo de Longpre, Wang e DuBois (2020) levanta a hipótese de que os métodos de AD aplicados ao PLN só podem ser benéficos se introduzirem novos padrões linguísticos. Desta forma, os autores sugerem que, ao considerar grandes modelos de língua pré-treinados, muitos métodos de AD não conseguem gerar maiores ganhos, pois tais modelos já são invariantes a diversas transformações.

Por isso, neste trabalho, investigaremos a hipótese de que o uso de determinados métodos de AD em PLN potencializa o treinamento dos modelos linguísticos no contexto de complexidade textual, especificamente na avaliação quantitativa da tarefa de AAL. Os métodos a serem explorados podem ser identificados na Tabela 2 por SS a nível da palavra e o de RT a nível de documento, ambos pertencentes ao espaço de dados. O motivo da escolha de tais métodos pode ser encontrada com mais detalhes no Capítulo 4.

[6]Não aplicável (N/A), uma vez que não necessariamente há controle na aplicação do método em termos de caractere, palavra, frase e/ou documento

2.3 Representação de Atributos

Conforme mencionado por Aggarwal (2018) e Freitas (2022), a alta disponibilidade de textos na *web*, redes sociais, correios eletrônicos, bibliotecas digitais e outros meios digitais nos últimos anos, tem potencializado o aumento de estudos em PLN orientado a dados. Segundo Freitas (2022), para a Associação para a Linguística Computacional, o campo de PLN ou linguística computacional busca fornecer modelos computacionais de fenômenos linguísticos baseados no conhecimento (*i.e.*, abordagens baseadas em regras) ou orientados por dados (*i.e.*, abordagens baseadas em AM).

No contexto de AAL, ainda que existam abordagens baseadas no conhecimento, tem se tornado cada vez mais comum o uso de modelos baseados em AM, vide tabela 1. Conforme explicado por Freitas (2022), o objetivo do AM é resolver problemas sem que os modelos tenham sido explicitamente programados para isto. Desta forma, as regras são aprendidas automaticamente por meio de exemplos, *i.e.*, dados de treino. No entanto, devido à natureza subjetiva e complexa da linguagem, que se manifesta em alta dimensionalidade e esparsidade (AGGARWAL, 2018), torna-se um desafio traduzir palavras em unidades discretas e estáveis (FREITAS, 2022). A representação de atributos de um texto é uma forma de comprimir dados, cuja natureza é de alta dimensionalidade, facilitando a identificação de padrões pela máquina (AGGARWAL, 2018).

Neste trabalho, explora-se tanto o uso das representações por incorporação de palavras (ou no inglês, *word embeddings* (MIKOLOV et al., 2013)), quanto dos conjuntos de métricas de análise linguística e psicolinguística disponíveis para o PB gerados pelo NILC-Metrix (LEAL; SCARTON et al., 2022), considerando análises combinatórias de ambos os formatos de representação de atributos. Ambas são explicadas em mais detalhes a seguir.

2.3.1 Vetorização de palavras

Um processo de vetorização de palavras ou *tokens*[7] busca representar palavras de um texto de forma numérica, *i.e.*, um vetor, a nível de palavra (AGGARWAL, 2018; FREITAS, 2022). Para distinção das abordagens de vetorização de palavras existentes atualmente, destaca-se uma das principais barreiras a este recurso mencionada por Freitas (2022): "o sentido não é uma propriedade intrínseca das palavras, mas uma abstração que só irá se concretizar no uso". Neste contexto, faz sentido segregar tais abordagens em dois grupos, conforme apresentado na Tabela 3.

Tabela 3: Abordagens para vetorização de palavras

Abordagem	**Definição**	**Modelos de referência**
Estática	Não pondera o uso da palavra corrente, *i.e.*, não considera o contexto durante a vetorização, e consequentemente é denominado estático. Uma abordagem estática presume que para cada palavra há apenas um conjunto de números para representá-la, e por isso, interpreta ambiguidades de modo equivalente.	*Bag-of-Words*, TF-IDF, Word2Vec, GloVe
Contextualizada	Segundo Freitas (2022), a abordagem contextualizada caracteriza o dinamismo próprio dos sentidos léxicos e por dependerem dos dados advindos de *corpora* atualizados consideram a linguagem como marcas de significado que se deslocam no espaço e tempo. De modo que, os vetores gerados são definidos para cada palavra em seu contexto corrente.	*Long Short-Term Memory* (LSTM) e *Transformers*

Considerando o exemplo apresentado por Freitas (2022), a palavra "rei" nas expressões *rei do gado*, *rei da cocada preta* e *rei da Espanha* seriam representadas por um mesmo vetor no caso de uma abordagem estática, enquanto que uma abordagem contextualizada consideraria a aplicação da palavra, gerando vetores diferentes.

No que dizem respeito aos modelos apresentados na Tabela 3, uma das tentativas de

[7]Segundo Freitas (2022), a tokenização é o processo de dividir um texto em unidades menores chamadas *tokens*. Tais unidades podem ser palavras, sinais de pontuação ou símbolos. Freitas (2022) aponta para três estratégias de tokenização: (1) baseada em palavra, que divide o texto em palavras gráficas, ou seja, o quê aparece entre espaços em branco ou sinais de pontuação; (2) baseada em caracteres, que cria uma representação numérica para cada letra; e (3) baseada em subpalavra, uma abordagem intermediária, que divide as palavras em pedaços menores mais informativos do que os caracteres mas com menos caracteres do que as palavras gráficas. A escolha da estratégia de tokenização depende do objetivo do modelo e do tamanho do texto. Ainda que conceitualmente *palavra* e *token* tenham significados diferentes, a primeira sendo uma unidade linguística e a segunda, uma unidade textual, dado o presente contexto de vetorização, ambos os termos serão utilizados de forma intercambiável neste trabalho.

codificar texto em formato numérico mais comum que conhecemos é a *bag-of-words*, ou em PB, saco de palavras (AGGARWAL, 2018). Nesta abordagem, a ordem das palavras não é relevante para codificação da informação, cada documento é representado por um vetor cuja dimensão corresponde ao número de palavras no texto e cujos valores correspondem à frequência de cada palavra, ou à presença ou não de uma palavra no texto em uma representação binária (AGGARWAL, 2018). O vetor de saída desta abordagem é classificado como esparso, isto significa que há uma grande quantidade de elemtentos com valor zero, ou, não presentes/necessários.

Conforme apontado por Aggarwal (2018), uma das principais questões relacionadas ao *bag-of-words* é que não há distinção de importância entre palavras. A abordagem que considera a normalização destas frequências e, consequentemente, pondera as ocorrências de palavras menos frequentes, é conhecida por *Term Frequency - Inverse Document Frequency (TF-IDF)*. Nesta abordagem, ao considerar uma coleção de textos $\{T_1, T_2, \ldots, T_i\}$, a frequência do termo (TF) da palavra é calculada com base no número de vezes que cada palavra aparece no texto T_i, enquanto a frequência do documento (DF) de uma palavra é calculada com base na razão entre a quantidade de textos da coleção que contêm pelo menos uma ocorrência da palavra, pelo número de textos total da coleção i. Em posse dos valores de TF e DF, cada palavra recebe uma pontuação dada pela multiplicação do TF pelo logaritmo do inverso do DF (AGGARWAL, 2018). Desta forma, se uma palavra é muito comum em todos os documentos, isto significa que não agrega muito valor para a análise, de modo que será representada com um TF-IDF mais baixo. O oposto também é válido. Apesar desta ponderação, destaca-se que este modelo também não considera a ordem das palavras dispostas no texto, em especial em modelos unigrama.

Diferentemente do *bag-of-words* e *TF-IDF*, que consideram métodos estatísticos para geração de vetores, o *word2vec* (MIKOLOV et al., 2013) faz uso de ML baseados em redes neurais (MA, 2022). Conforme apontado na Tabela 2, um ML consiste em um modelo probabilístico cujo intuito é prever palavras dado um determinado contexto, ou vizinhança de palavras. No entanto, ressalta-se que o objetivo do *word2vec* não é prever a próxima palavra em uma sequência como usualmente é realizado nos MLs, mas sim representar palavras como vetores densos em um espaço n-dimensional dadas pela probabilidade das palavras em seu contexto. Assim, o *word2vec* pode ser treinado de duas formas distintas, conforme indicado por Aggarwal (2018):

- *Continuous Bag of Words (CBOW)*: ao considerar uma rede neural, pode-se dizer que nesta abordagem, a entrada é uma representação média da vizinhança da palavra

que se deseja prever; e a saída é a probabilidade de cada uma das palavras do vocabulário aprendido ser a palavra que se deseja prever.

- *Skip-Gram (SG)*: neste caso, a entrada é a representação binária de uma palavra central; e a saída é um vetor de probabilidade das palavras vizinhas dado o vocabulário aprendido pelo modelo.

Independente da abordagem, os pesos das camadas ocultas da rede neural do *word2vec* compõem o vetor de incorporação de palavras (AGGARWAL, 2018). Esta incorporação tende a posicionar palavras que são vizinhas nos textos de treino, dada uma janela de n palavras, mais próximas umas das outras no espaço latente[8] (FREITAS, 2022; AGGARWAL, 2018). Isto permite, portanto, capturar características semântico-lexicais e transportá-las para um espaço vetorial. Sendo assim, enquanto *bag-of-words* e *TF-IDF* geram vetores esparsos para representar as palavras, o *word2vec* gera vetores densos cujas dimensões são fixas e parametrizáveis (dada pelo número de neurônios da camada intermediária da rede neural) (JURASFKY; MARTIN, 2020). Além disso, enquanto os métodos estatísticos assumem que palavras como "rei" e "monarca" são representadas como vetores distintos, a representação por incorporação é capaz de captar a relação de similaridade entre essas duas palavras (a depender dos dados de treino do modelo) e gerar vetores mais próximos entre si (JURASFKY; MARTIN, 2020).

No que tange o PB, destacam-se duas bibliotecas comumente utilizadas de código aberto que fazem uso do *word2vec*:

1. *Fasttext* (INC, 2022): a biblioteca foi criada pelo laboratório de pesquisa de Inteligência Artificial (IA) do *Facebook* e tem integração para carregamento dos modelos pré-treinados apresentados por Edouard et al. (2018). Ainda que o *Fasttext* tenha sido desenvolvido com base no *word2vec*, o *Fasttext* considera em seu treinamento n-gramas de caracteres (ou seja, subpalavras), ao invés de *tokens*, como é o padrão do *word2vec*. A incorporação de informações de subpalavras tem por objetivo lidar com palavras fora do vocabulário (*Out of Vocabulary* - OOV), uma vez que permite que o modelo represente palavras que não haviam sido vistas durante o treinamento por meio da combinação das representações de suas subpalavras. Assim, os modelos *Fasttext* foram treinados considerando sequências de cinco caracteres, usando

[8] Um espaço latente pode ser definido como um espaço abstrato multi-dimensional que codifica uma representação interna significativa de eventos observados externamente (CHAQUET-ULLDEMOLINS et al., 2022).

CBOW com número de dimensões (ou número parametrizado de neurônios) em 300, e fazendo uso da técnica de *negative sampling*, ou em PB, amostragem negativa[9].

2. *Gensim* (ŘEHŮŘEK, 2022): acrônimo para *Generate Similar*, a biblioteca começou em 2008 com uma coleção de *scripts* em *Python* para um projeto da Biblioteca Checa de Matemática Digital. Atualmente, conta com diversos algoritmos não supervisionados, como *word2vec*, *fasttext*, *latent semantic indexing* e outros. Esta biblioteca permite carregamento dos modelos pré-treinados no trabalho desenvolvido por Hartmann, Fonseca et al. (2017) e disponibilizados no repositório do NILC (LINGUÍSTICA COMPUTACIONAL, 2017). O repositório conta com modelos baseados em *word2vec*, *fasttext*, *glove* e *wang2vec* com as variações *CBOW* e *skip-gram* cujos vetores de palavras foram gerados em várias dimensões.

Conforme elucidado por Ma (2022), no *word2vec*, as incorporações de palavras são adquiridas por um modelo não supervisionado cuja função objetivo é maximizada pela probabilidade condicional de coocorrência de palavras que frequentemente coocorrem no conjunto de treino. Já o *GloVe* considera como parte da função objetivo informações de frequência global do conjunto de treino. Ainda que tais modelos considerem o contexto das palavras no momento do treino, eles geram apenas uma representação vetorial para cada palavra, e por isso, se enquadram como estáticos na Tabela 3. Os modelos contextualizados apresentados na Tabela 3 também fazem uso de redes neurais. Entretanto, consideram a aplicação de redes mais profundas, *i.e.*, com várias camadas ocultas, com criação das incorporações de palavras de modo dinâmico.

O modelo *ELMo* (PETERS et al., 2018), acrônimo para *Embeddings from Language Model*, usa um arquitetura bidirecional profunda de LSTM (HOCHREITER; SCHMIDHUBER, 1997), ou em PB, memória de curto e longo prazo. A rede LSTM é um tipo especial de rede neural recorrente (em inglês RNN), que permite o uso de recorrências circulares, *i.e.*, *loops*, para modelar dependências de linguagem (AGGARWAL, 2018). Conforme apresentado por Aggarwal (2018), as RNNs consideram o uso da informação sequencial, *i.e.*, a ordem de disposição das palavras no texto, de modo que a saída para cada elemento depende dos cálculos anteriores. No entanto, as RNNs tradicionais, conhecidas no inglês por *vanilla*, não conseguem armazenar informações muito distantes em um

[9] Ao invés de tentar prever a probabilidade para todas as palavras do vocabulário aprendido, a técnica de *negative sampling* tenta prever a probabilidade para um conjunto aleatório reduzido de palavras que não pertencem ao texto em questão (AGGARWAL, 2018). Por isso, a amostragem negativa permite alterar apenas parte dos pesos da camada de incorporação, ao invés de todos para cada amostra do treinamento do modelo, proporcionando um processamento mais eficiente do modelo (AGGARWAL, 2018).

texto devido ao desaparecimento do gradiente probabilístico a medida que o texto evolui. Por exemplo, suponha que o modelo precise prever a palavra sublinhada no seguinte trecho: "O Brasil é minha terra natal, com o tempo fui morar na Inglaterra, e hoje sei falar inglês. Também sei (...), mas a minha língua nativa é o português", uma *RNN vanilla* não conseguiria correlacionar o início do texto com o final por conta da distância entre as palavras. No caso das LSTM, é possível modificar as condições de recorrências de como os estados ocultos do modelo são propagados (AGGARWAL, 2018). Assim, a LSTM cria uma espécie de célula de memória de longo prazo capaz de persistir informações.

Os modelos conhecidos por *Transformer* processam uma sequência de palavras de uma só vez e mapeiam as dependências relevantes entre as palavras, independente de quão distantes as palavras aparecem no texto (VASWANI et al., 2017). Conforme apontado por Ma (2022), os *Transformers* são o estado-da-arte em modelos neurais de linguagem e têm uma vantagem em relação ao LSTM e RNN porque suas camadas são independentes e podem ser processadas em paralelo. Neste sentido, apesar das RNNs apresentarem menor demanda computacional do que as redes LSTM e, por sua vez, a LSTM apresentar menor demanda que a arquitetura *Transformer*; em geral, os modelos *Transformer* conseguem processar sequências de texto mais rapidamente que as redes LSTM e RNN devido à alta capacidade de paralelização durante o treinamento[10].

Os modelos *Transformers* são baseados em uma arquitetura codificador-decodificador com uso de mecanismos de atenção (VASWANI et al., 2017). Em outras palavras, a arquitetura *Transformer* é composta por camadas de entrada, que transformam os dados de entrada em uma representação vetorial; e camadas de saída, capazes de transformar essas representações em saídas legíveis. Entre essas camadas, existem camadas de atenção, que calculam a importância de cada elemento em relação aos outros elementos do conjunto de dados por meio do acesso aos estados ocultos da rede. Assim, a rede é capaz de capturar dependências entre as palavras.

O *Generative Pre-trained Transformer* (GPT) é um modelo *Transformer*-decodificador, *i.e.*, a arquitetura do GPT é baseada em um modelo *Transformer* com apenas um bloco de codificador (ou *encoder*, em inglês), e vários blocos do decodificador (ou *decoder*, em inglês). O modelo é treinado previamente com um conjunto de dados com grandes quantidades de textos, de modo que, é capaz de gerar textos de forma autônoma e pode ser ajustado para desempenhar tarefas específicas por meio do método de *fine-tuning* (RADFORD et al., 2018). O *fine-tuning* permite que uma rede treinada previamente

[10]Destaca-se, no entanto, que a complexidade computacional exata depende do tamanho do modelo, do tamanho do conjunto de dados e dos recursos de *hardware* disponíveis.

seja ajustada para desempenhar melhor uma tarefa específica, sem ter que treiná-la por completo novamente.

Segundo Devlin et al. (2019), um dos gargalos relacionados tanto ao *Elmo* quanto ao GPT é a unidirecionalidade do modelo para aprender representações gerais de linguagem. Isto significa que a saída destes modelos para cada posição depende apenas das palavras anteriores na sequência. Neste contexto, Devlin et al. (2019) propuseram o *Bidirectional Encoder Representations from Transformers* (BERT), que faz uso de um modelo de linguagem mascarado inspirado no *cloze procedure*, mencionado na Tabela 1.

Diferentemente do GPT, o BERT não usa blocos de decodificação em sua arquitetura. O modelo é baseado em uma arquitetura de codificador, em que a entrada do *token* é processada por múltiplas camadas de codificação, e cada camada é responsável por extrair informações de alto nível a partir dos dados de entrada. Ainda que estruturalmente *GPT* e *BERT* se distanciem entre si, ambos são modelos de linguagem muito poderosos, que têm sido aperfeiçoados ao longo do tempo e inspirado diversos trabalhos. O BERT atende muito bem tarefas supervisionadas de PLN, e o GPT tarefas não supervisionadas de geração de texto. Um ponto de atenção é que estas redes neurais são computacionalmente intensas e precisam de muitos recursos para treino. Isto significa que um número elevado de *tokens* processados ao mesmo tempo podem levar a resultados imprecisos ou impraticáveis (SOUZA; NOGUEIRA; LOTUFO, 2020). Por isso, essas abordagens são projetadas para lidar com textos de tamanho moderado, e tendem a ter restrições no número de *tokens* a serem processados.

O BERTimbau é um modelo BERT pré-treinado especificamente para o PB (SOUZA; NOGUEIRA; LOTUFO, 2020), e o modo mais comum de carregá-lo localmente é por meio da biblioteca de código aberto *transformers*, disponibilizada pelo grupo *Hugging Face* (WOLF et al., 2020).

2.3.2 Métricas de análise linguística e psicolinguística

Métricas de análise linguística e psicolinguística são embasadas em medidas textuais de coesão, coerência, nível de complexidade, clareza, precisão e consistência. Cada uma dessas métricas mede aspectos da qualidade de um texto e pode ser usada para avaliar textos em diferentes domínios. Isto significa que há uma relação entre leiturabilidade e tais métricas. Conforme supracitado na Tabela 1, a ferramenta NILC-Metrix (LEAL; SANCHES DURAN et al., 2021; LEAL; SCARTON et al., 2022) é um dos recursos mais completos capaz de gerar métricas de análise linguística e psicolinguística para o PB. As

métricas disponíveis no NILC-Metrix estão divididas em 14 grupos (LEAL; SANCHES DURAN et al., 2021):

1. Medidas Descritivas (dez métricas): descreve estatísticas básicas de um texto, como número de palavras, de parágrafos, de sentenças e outros.

2. Simplicidade Textual (oito métricas): avalia o quão fácil é um texto por meio de proporções, como a proporção pela quantidade de sentenças longas, de conjunções difíceis, de pronomes de proximidade e outros.

3. Coesão Referencial (nove métricas): captura a presença de elementos necessários para relacionar sentenças.

4. Coesão Semântica (11 métricas): considera a sobreposição de palavras semanticamente relacionadas por meio do método de Análise Semântica Latente (LSA) (LANDAUER et al., 1997). O modelo LSA para NILC-Metrix foi treinado com o *corpus* BrWaC (WAGNER FILHO; WILKENS; IDIART et al., 2018), que contém 300 dimensões.

5. Medidas Psicolinguísticas (24 métricas): relacionadas à facilidade do texto em termos de idade de aquisição das palavras, concretude, familiaridade e imageabilidade das palavras. O recurso lexical utilizado pelas métricas desse conjunto contém 26.874 palavras (palavras de conteúdo), de modo que se uma palavra do texto não for incluída no recurso, essas métricas são afetadas.

6. Diversidade Lexical (15 métricas): obtida através da Razão Tipo-*Token* (TTR), ou seja, o número de palavras de um determinado tipo desconsiderando repetições (de modo que, um tipo que pode ser um verbo, substantivo, etc.) dividido pelo número de *tokens* (todas as palavras do texto, considerando repetições). A diversidade lexical é inversamente proporcional à coesão: quanto menor a diversidade lexical, maior a coesão.

7. Conectivos (12 métricas): estabelece proporções dos conectivos no texto, bem de quatro tipos diferentes de conectivos: aditivo, causal, lógico e temporal.

8. Léxico Temporal (12 métricas): detalha as ocorrências relativas de cada tempo verbal e modo em relação ao total de tempos e modos verbais no texto.

9. Complexidade Sintática (27 métricas): utiliza dados de árvores de dependência, que consideram características sintáticas em memória, como o número de palavras antes

do verbo principal de uma sentença. Este agrupamento considera os índices propostos por Yngve (1960) e Frazier (1985), bem como diversas medidas de proporção envolvendo orações.

10. Densidade de Padrões Sintáticos (quatro métricas): apresenta métricas correlacionadas à dificuldade de processamento do texto: orações em gerúndio, número médio de palavras por sintagma nominal[11], número máximo e mínimo de palavras por sintagma nominal.

11. Informações Morfossintáticas de Palavras (42 métricas): calcula medidas tradicionais de densidades de conteúdo e de palavras funcionais, no texto e por frase, bem como uma série de desdobramentos dessas densidades, dadas por: adjetivos, advérbios, verbos, substantivos, preposições e pronomes.

12. Informações Semânticas de Palavras (11 métricas): avalia proporções de palavras com polaridade negativa/positiva em relação a todas as palavras do texto, bem como medidas de ambiguidade (de palavras de conteúdo, e em detalhe por substantivos, adjetivos, verbos e advérbios) e de métricas relacionadas à proporção de substantivos abstratos e nomes próprios nas frases e no texto.

13. Frequência de Palavras (dez métricas): considera frequências de todas as palavras do conteúdo e das palavras mais raras do texto a partir dos conjuntos: Banco de Português (DAVIES, 2017), *Corpus* Brasileiro (SARDINHA, 2010) e do BrWaC (WAGNER FILHO; WILKENS; IDIART et al., 2018) com normalizações usando frequência por milhão e escala logarítmica zipf [12].

14. Índices de Leiturabilidade (cinco métricas): reúne cinco fórmulas clássicas usadas para avaliar a legibilidade do texto: Brunet (THOMAS et al., 2005), Dale Chall (DALE; CHALL, 1948), Flesch (KINCAID; FISHBURNE JR et al., 1975), Gunning's Fog (BROWN, 1997) e Honore (THOMAS et al., 2005).

Observa-se, portanto, que no contexto deste trabalho, faz sentido considerar as métricas textuais supracitadas como entrada para modelos de AAL.

[11] Palavra ou conjunto de palavras que têm como núcleo o substantivo (KURAMOTO, 1996).

[12] A frequência na escala Zipf é calculada como $log_{10}(x)+3$, tal que x é o valor da frequência normalizada (LEAL; SCARTON et al., 2022).

3 Trabalhos Relacionados

Conforme elucidado por Imperial (2021), a AAL é a tarefa de avaliar o nível de facilidade ou dificuldade de compreensão de documentos de texto. No contexto de AM, é mais frequentemente vista como uma tarefa de classificação onde um conjunto de textos é treinado com seus rótulos correspondentes. No que diz respeito ao estado-da-arte em AAL, destacam-se os modelos propostos por Martinc, Pollak e Robnik-Šikonja (2021) e Lee, Jang e Lee (2021).

Martinc, Pollak e Robnik-Šikonja (2021) foram capazes de aumentar a precisão da classificação em um *corpus* de língua inglesa, o *Weebit* (FENG; ELHADAD; HUENER-FAUTH, 2009), por meio de um modelo computacional de língua que amplia em cerca de 4% a precisão por meio da incorporação de atributos contextualizados. De acordo com Lee, Jang e Lee (2021), este resultado de incremento sugeriu, de forma inédita, que modelos de redes neurais com atributos gerados automaticamente por transferência de aprendizado podem ser mais eficazes do que os modelos de AM tradicionais na tarefa de AAL. No trabalho de Lee, Jang e Lee (2021), além do *corpus Weebit*, são considerados outros dois *corpora* da língua inglesa, *OneStopEnglish* e *Cambridge* (XIA; KOCHMAR; BRISCOE, 2019). Os autores exploram o uso de um modelo híbrido, em que são combinados os resultados das previsões de um classificador por transferência de aprendizado, *i.e.*, por meio de uma rede neural considerando um modelo de incorporação de palavras pré-treinado, e atributos linguísticos envolvidos por um modelo não-neural, como um Regressor Logístico (RL). Em comparação com o estudo anterior de Martinc, Pollak e Robnik-Šikonja (2021), Lee, Jang e Lee (2021) conseguiram um aumento de aproximadamente 20%, atingindo uma acurácia de 99% para o *corpus OneStopEnglish*. Este valor notório é apontado como uma preocupação de *overfitting* para aquele domínio textual.

Assim sendo, segundo os autores antes citados, sem uma metodologia capaz de conectar vários conjuntos de dados ou um novo grande conjunto público de dados para AAL, será sempre um desafio desenvolver um modelo computacional de língua de uso geral. Adicionalmente, uma conclusão importante na análise de Lee, Jang e Lee (2021)

é que *conjuntos menores de dados se beneficiam mais do uso de atributos linguísticos do que da incorporação de palavras*. Isto significa que o uso do classificador por rede neural pré-treinado aumentou ainda mais o *overfitting* para o *corpus Cambridge*[1].

Ainda que os primeiros trabalhos de AAL já tenham pouco mais de duas décadas de existência (vide Tabela 1), trabalhos voltados para o PB foram fomentados a partir do surgimento do Coh-Metrix-PORT (MCNAMARA DS.; GRAESSER, 2002). No entanto, conforme apontado por Scarton, Gasperin e Aluisio (2010), até o ano de 2010, a única ferramenta a considerar abordagens em PLN para avaliação de leiturabilidade no PB havia sido realizada pelo trabalho de Scarton, Gasperin e Aluisio (2010).

Em trabalhos que lidam com outros idiomas com menos recursos do que a língua inglesa, como é o nosso caso com o PB, destaca-se o estudo feito por Imperial (2021) para o idioma filipino. Segundo o autor, evidências de eficácia de métodos baseados em transferência de aprendizado, como no trabalho de Martinc, Pollak e Robnik-Šikonja (2021), valem apenas para conjuntos de dados com alta volumetria. Assim, Imperial (2021) propõe um método que combina o uso de métodos tradicionais de aprendizado, como RL e SVM, com atributos de incorporação de palavras oriundos do modelo BERT e métricas linguísticas concatenados. Foi demonstrado que o conhecimento implicitamente codificado na incorporação de palavras contextualizada pode ser usado como um conjunto de recursos completos para idiomas com recursos de baixa volumetria. Para o *corpus Adarna House* que contém 265 livros em filipino, o uso da concatenação dos atributos superou o uso individual destes conjuntos de atributos, de modo a alcançar um *F1-Score* de 0.571 para o classificador SVM.

Neste contexto, entende-se que mesmo que a leiturabilidade apresente características que possam ser analisadas de maneira global, cada idioma apresenta sua particularidades sintáticas e semânticas, bem como limitações de recursos e ferramentas, de modo que faz sentido apresentar nesta seção, a aplicação da tarefa de AAL dos trabalhos em PB. Assim sendo, foi elaborada a Tabela 4 com um breve descritivo de todos os trabalhos identificados pela autora envolvendo a temática em questão e baseados em AM especificamente para o PB. Cada trabalho da Tabela 4 foi representado por um índice dado pela coluna *n*.

[1] Apenas para fins comparativos, o *corpus Weebit* contém 3125 textos, o *OneStopEnglish*, 567, e o *Cambridge*, 331.

Tabela 4: Trabalhos de avaliação automática de leiturabilidade para o PB

n	Referência(s)	Resumo
1	Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português (SCARTON; ALUÍSIO, 2010)	Este artigo apresenta a primeira adaptação da ferramenta Coh-Metrix para o PB e duas aplicações. Dentre as aplicações apresentadas, é considerada a criação de classificadores binários entre textos complexos e simples. Para isto, foi desenvolvido um *corpus* combinado, formado por textos classificados conforme o público alvo da revista ou jornal de origem. O melhor classificador treinado foi aquele implementado com SVM do WEKA e atingiu em média um *F-score* de 0.97, considerando as métricas do Coh-Metrix-Port combinadas com o índice Flesch como sendo os atributos de entrada do modelo.
2	*Readability Assessment for Text Simplification* (ALUISIO et al., 2010)	Neste trabalho, Aluisio et al. (2010) apresentam um recurso incorporado à ferramenta Simplifica (SCARTON; OLIVEIRA et al., 2010)(mencionada na Tabela 1), cujo intuito é o de categorizar textos em três níveis de leiturabilidade: (1) textos originais voltados para leitores avançados, (2) textos naturalmente simplificados voltados para pessoas com nível de alfabetização básico e (3) textos fortemente simplificados para pessoas com nível de alfabetização rudimentar. Como atributos do modelo classificatório, foram considerados três grupos de atributos: o primeiro contém atributos derivados da ferramenta Coh-Metrix-PORT, o segundo contém características que refletem a incidência de construções sintáticas particulares desenvolvidas pelos próprios autores, e o terceiro contém características derivadas de um modelo de língua estatístico desenvolvido com base na ferramenta SRILM (STOLCKE, 2002), produzido com artigos advindos da *Folha de São Paulo*.

Tabela 4: Trabalhos de avaliação automática de leiturabilidade para o PB (Continuação)

n	Referência(s)	Resumo
		O modelo classificatório com melhor resultado foi aquele que considerou a combinação dos três grupos de atributos e treinados por SVM, de modo a alcançar um *F-Score* de 0.913 para o nível (1), 0.483 para (2) e 0.732 para (3).
3	*Revisiting the Readability Assessment of Texts in Portuguese*(SCARTON; GASPERIN; ALUISIO, 2010)	Scarton, Gasperin e Aluisio (2010) consideram duas classes para distinção dos textos entre simples (para leitores entre 7 e 14 anos) e complexo (para adultos), usando o Coh-Metrix-Port com 40 métricas textuais. Nos testes, avaliou-se o impacto de diferentes gêneros e domínios por meio da criação de dois classificadores independentes, um treinado apenas com textos de notícias e outro treinado apenas com textos de divulgação científica. Adicionalmente, experimentaram-se algoritmos de seleção de atributos, a fim de selecionar os atributos mais relevantes de um conjunto extraído do Coh-Metrix Port. Os experimentos mostraram que o algoritmo SVM, com todas as métricas do Coh-Metrix Port, obteve o melhor desempenho e que o classificador treinado em textos de jornal foi capaz de generalizar e classificar consideravelmente bem os textos de divulgação científica.
4	*Crawling by Readability Level* (WILKENS; ZILIO et al., 2016)	Wilkens, Zilio et al. (2016) fizeram uso da iniciativa *Web as Corpus* (WaC) [2] como uma ferramenta para geração de grandes *corpora* classificados em termos de leiturabilidade. Para anotação dos *corpora* coletados, foi utilizado um modelo de regressão estatístico cujos atributos de entrada foram sete métricas de coesão e coerência textual com baixo custo computacional, geradas pelos próprios autores.

[2]WaC é um conjunto de ferramentas que permite o acesso à *World Wide Web* como um corpus.

Tabela 4: Trabalhos de avaliação automática de leiturabilidade para o PB (Continuação)

n	Referência(s)	Resumo
		A entrada do modelo foi dada pelo *corpus* mais controlado, o *Wikibooks*[3], que é separado em três níveis do sistema educacional brasileiro: 33 livros do ensino fundamental; 65 livros do ensino médio e 21 livros de graduação. E, o teste foi aplicado em dois sentidos: (1) considerando o próprio *Wikibooks*, em que o modelo delineado atingiu um *F-score* médio de 0.691 para as três classes; (2) considerando o WaC, em que não foi possível definir uma taxa de acurácia uma vez que se trata de um *corpus* não classificado, mas foi realizada uma análise estatística preliminar a partir de propriedades distribucionais, lexicais e sintáticas em comparação às duas classes do Wikibooks.
5	*Automatic construction of large readability corpora* (WAGNER FILHO; WILKENS; VILLAVICENCIO, 2016)	Os autores apresentam uma continuação do trabalho Wilkens, Zilio et al. (2016), criando um *corpus* de trabalho tanto com o PB quanto com o inglês. Neste artigo, os autores consideram outras representações de atributos, que categorizam em três grupos conforme o custo computacional: (1) superficial (contagens e listas), (2) médio (dependente das marcações de partes, *i.e., POS taggers*) e (3) profundo (dependente de um *parser* ou de uma *WordNet*). Adicionalmente, consideraram outros *corpora* além do *Wikibooks* usado no trabalho anterior. Os autores concluem que os atributos superficiais apresentam boa performance de classificação em textos em português, e principalmente para a língua inglesa. No entanto, o uso completo dos três grupos de atributos melhora os resultados para a maioria dos *corpora*.

[3] https://pt.wikibooks.org/

Tabela 4: Trabalhos de avaliação automática de leiturabilidade para o PB (Continuação)

n	Referência(s)	Resumo
		No caso, do *Wikibooks*, o melhor resultado foi por uso apenas dos atributos superficiais, atingindo um *F-Score* médio de 0.75. Segundo Wagner Filho, Wilkens e Villavicencio (2016) o RL produziu os melhores resultados entre os classificadores testados, sendo melhores para classes binárias do que com variações em classes de leiturabilidade.
6	Predição da complexidade sentencial de recursos educacionais abertos em português (GAZZOLA; LEAL; ALUISIO, 2019)	Este trabalho explora o uso de quatro classes de leiturabilidade, com base nos níveis do sistema educacional brasileiro, usando *corpus* de gêneros de texto distintos com 2067 extratos. Gazzola, Leal e Aluisio (2019) trazem uma avaliação com RL, SVM, *Random-Forest* e *Multilayer Perception*, considerando 79 métricas textuais com aplicação de técnicas distintas para seleção de atributos. O melhor modelo treinado foi com o SVM e atingiu uma média ponderada de *F-Score* de 0.804. Identificou-se que o desempenho dos modelos com seleção de atributos foi inferior ao do modelo com todos os atributos.
7	Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular (LEAL, 2019)	O foco deste trabalho é na avaliação de métodos de predição de complexidade de frases para o PB. Para isto, foram criados dois *corpora* de sentenças, o *PorSimplesSent* (LEAL; DURAN; ALUÍSIO, 2018), e outro com métricas de rastreamento ocular e normas de previsibilidade para estudantes de nível superior, denominado *RastrOS* (LEAL; LUKASOVA et al., 2022). Foi considerada a versão mais recente da ferramenta NILC-Metrix (LEAL; SANCHES DURAN et al., 2021) (com 200 métricas), bem como abordagens de transferência de aprendizado com adição das métricas de rastreamento ocular.

Tabela 4: Trabalhos de avaliação automática de leiturabilidade para o PB (Continuação)

n	Referência(s)	Resumo
		Leal (2019) atingiu o nível do estado-da-arte para a tarefa de predição da complexidade de frases no PB, com 97,5% de acurácia. Com base no melhor método desenvolvido, o autor criou a aplicação *Simpligo* (LEAL; MAGALHAES et al., 2019; LEAL, 2020), que atribui um índice de complexidade individual para a sentença informada.
8	Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental(HARTMANN; ALUÍSIO, 2020)	Ainda que este trabalho não esteja direcionado diretamente para a tarefa de AAL, ao investigar etapas do processo de adaptação textual, são consideradas abordagens para a identificação de palavras complexas. No processo de classificação de palavras em simples ou complexas, Hartmann e Aluísio (2020) avaliaram desde abordagens clássicas até as mais modernas, incluindo uso de incorporação de palavras contextualizadas. O uso da incorporação de palavras oriunda do Elmo apresentou melhor desempenho com acurácia média de 97,7%.

Como forma de apresentar uma visão mais objetiva dos trabalhos voltados especificamente para a análise de textos mencionados na Tabela 4, propôs-se a Tabela 5. Neste caso, os trabalhos são representados apenas pelos índices dados pela coluna *n* da Tabela 4. Sobre a metodologia utilizada para classificação dos *corpora*, todos os estudos apresentados na Tabela 5 realizaram uma atribuição manual em que a cada *corpus* foi atribuída uma classe conforme seu público-alvo. No caso dos estudos 2 e 5, os *corpora* selecionados foram baseados em textos do tipo "complexo", de modo que foram criadas versões simplificadas dos textos por linguistas e incluídas aos *corpora* para balanceamento[4].

Para facilitar o dimensionamento da Tabela 5, foram utilizadas as seguintes siglas:

- BRE: BrEscola
- CC: Caderno Ciência
- CH: Ciência Hoje
- CHC: Ciência Hoje das Crianças

[4]Com exceção para o *corpus* do *Wikibooks* no estudo 5, em que a classe atribuída foi pelo nível escolar.

Tabela 5: Resumo comparativo dos principais modelos de AAL para o PB

n	CL	Corpora	QT	QTP	DD	SA	MAMR	FS
1	2	PSFL + CC + JCBC + ZH + CHC + CH	959	265.862	UI	41 do CMP	SVM	0.97
2	3	CC + ZZ + CH	592	297.674	Não	42 do CMP + 6 por EP + 10 por ML	SVM	0.71
3	2	ZH + CH + PSFL + CHC + CC + DG	689	241.500	UI	48 do CMP	SVM	0.94
4	3	WB	77	636.309	Sim	7 por EP	MR	0.69
5	2-3	WB + ESC + PSFL + ZH + BRE	9.829	4.750.690	UI	134 por EP	RL	0.75
6	4	PSFL + ZH + ES + WB + EE	2.067	813.417	Sim	79 do CMP	SVM	0.80

- CL: Número de Classes de Leiturabilidade
- CMP: Coh-Metrix-Port
- DD: Disponível para *Download*
- DG: Diário Gaúcho
- EE: Exames do Enem
- EP: Elaboração Própria
- ES: Exames do SAEB
- ESC: É Só o Começo
- FS: *F1-Score* Médio
- JCBC: Jornal da Cidade de Bauru (Criança)
- MAMR: Modelo de Aprendizado com Melhor Resultado
- MR: Modelo de Regressão
- PSFL: Para o Seu Filho Ler
- QT: Quantidade de textos
- QTP: Quantidade de tokens/palavras
- SA: Seleção de Atributos
- UI: URL inacessível
- WB: Wikibooks
- ZH: Zero Hora

Certamente, o grupo de pesquisa NILC é uma referência no uso de PLN avançado em diversas tarefas, incluindo a de ST e de análises para níveis linguísticos variados, dedicando-se especialmente ao PB. No contexto de trabalhos sobre leiturabilidade, destaca-se, no NILC, o projeto PorSimples (ALUÍSIO et al., 2008). Mais recentemente, o NILC desenvolveu o trabalho de Leal (2019) sobre medidas de complexidade textual, o que culminou na criação da ferramenta *Simpligo* (LEAL; MAGALHAES et al., 2019)[5].

No entanto, a tarefa aqui proposta se diferencia de tais trabalhos ao se concentrar na classificação textual para estimar a leiturabilidade e no uso de ferramentas que sejam capazes de mitigar o gargalo da falta de grandes quantidades de dados linguísticos. Assim, aproxima-se mais do trabalho proposto por Wilkens, Zilio et al. (2016) e Imperial (2021). O trabalho de Wilkens, Zilio et al. (2016) é particularmente interessante haja vista que os autores visam a classificação por leiturabilidade e sugerem o uso da iniciativa WaC para reduzir o gargalo linguístico. E, o trabalho de Imperial (2021) busca utilizar recursos disponíveis em idiomas de baixo recurso para potencializar o AM na tarefa de AAL. Entretanto, nenhum deles propõe o uso de métodos de AD.

[5]O Simpligo se encontra disponível para uso *on-line* na plataforma do NILC(LEAL, 2020).

4 Aumento de Dados (AD) para a tarefa de Avaliação Automática de Leiturabilidade (AAL)

Para criação dos métodos de AD propostos neste trabalho, consideraram-se tarefas agnósticas[1] de AD devido à facilidade de implementação e menor custo computacional. Quanto aos métodos de AD selecionados, é importante entender a sua escolha conforme o contexto deste trabalho. Esse contexto pode ser resumido em duas premissas:

- Os recursos de PLN disponíveis para a língua portuguesa (ainda que não tão abrangentes quanto da língua inglesa), vide Tabela 1, permitem o uso de mecanismos para substituições e conversões interlinguísticas;
- Conforme Finatto (2020), a condição de leiturabilidade de um texto é multifatorial. Está fortemente atrelada ao tema e ao estilo do texto, bem como ao uso de terminologias, de vocabulário mais ou menos frequente, tipo de organização sintática ou de tipos de frases, entre outros tantos elementos.

Sendo assim, métodos de AD que envolvessem maior perda potencial de significado (seja por redução de contexto ou inclusão de ruídos) foram desconsiderados nesta análise. Com esta restrição, foram selecionados e desenvolvidos métodos nas seguintes categorias:

1. Substituição por Sinônimo (SS): é um método bastante popular de AD que parafraseia instâncias de texto substituindo certas palavras por sinônimos. Esta tarefa se aproxima da Simplificação Léxica (SL), no entanto, a SL pertence à área de adaptação textual e tem por finalidade reduzir a complexidade lexical ou sintática de um texto, preservando seu significado (HARTMANN; ALUÍSIO, 2020). Assim, estas tarefas se assemelham uma vez que estão relacionadas a mudanças adaptadas

[1] Tarefas que são generalizáveis e consequentemente, não são cunhadas para nenhuma tarefa em particular (LONGPRE; WANG; DUBOIS, 2020).

do vocabulário das sentenças. Todavia, conforme pontuado por Wilkens, Vecchia et al. (2014), a SL tem por objetivo substituir palavras complexas por sinônimos ou palavras semanticamente próximas, que sejam de mais fácil compreensão. Logo, a diferença entre SS e SL reside no fato de que, em termos semânticos, é possível alterar palavras e/ou expressões por um outro conjunto de palavras que não necessariamente estabeleçam uma relação de sinonímia entre si.

2. Retrotradução (RT), ou do inglês *Backtranslation* (BT): esta abordagem é particularmente interessante porque possui alta capacidade de parafraseamento, permitindo alterações tanto léxicas[2] quanto sintáticas[3] (BAYER; KAUFHOLD; REUTER, 2021). Segundo Bayer, Kaufhold e Reuter (2021), por se utilizar a tarefa de tradução do texto, o conteúdo é preservado e apenas as características estilísticas baseadas nos traços do autor são excluídas ou alteradas.

A seguir, explica-se como foram desenvolvidos os métodos de AD propostos neste estudo, antes citados, considerando uma argumentação a partir de exemplos de referência.

4.1 Substituição por Sinônimo (SS)

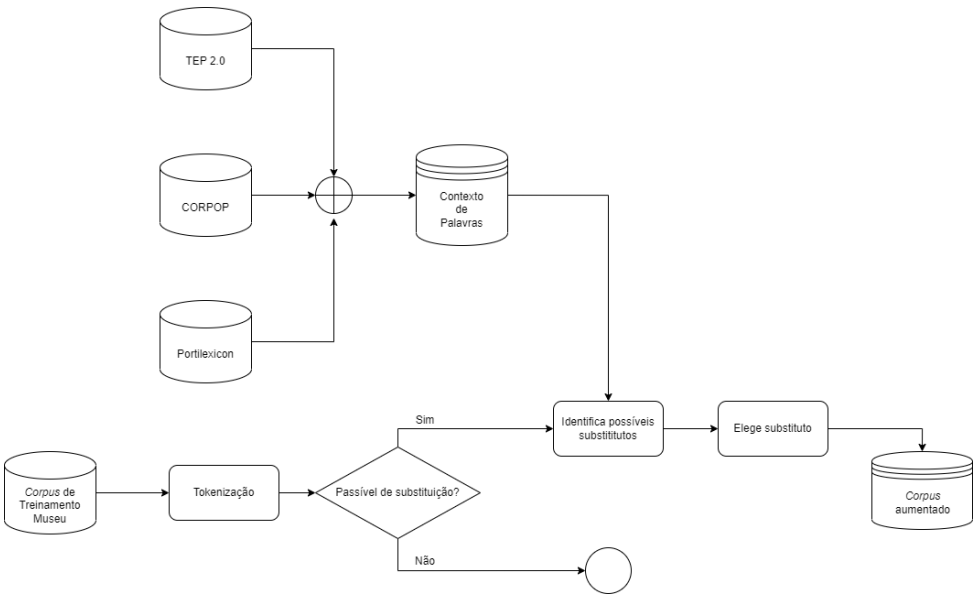


Figura 2: Fluxograma do Método - Substituição por Sinônimo (SS)

Para a execução desta tarefa de substituição, foi necessário inicialmente criar uma lista guia para o contexto de palavras de modo a identificar aquelas que estabelecessem

[2]O processo de simplificação léxico examina a escolha do vocabulário das sentenças.
[3]O processo de simplificação sintático avalia a disposição e relação lógica das palavras.

uma relação de sinonímia entre si. Para isto, partiu-se da definição de um universo de palavras que são simples, diferentemente do formato de substituição apresentado por Wilkens, Vecchia et al. (2014), em que os autores definem como primeira etapa a definição de palavras consideradas difíceis para serem substituídas.

Para construir este contexto de palavras, foram considerados os seguintes conjuntos de dados disponíveis para o PB:

- Tep 2.0 (DIAS-DA-SILVA; MORAES, 2003): uma coleção de palavras do PB e respectivos sinônimos, agrupadas em conjuntos (USP, N., 2022); no inglês este tipo de coleção é conhecido por *synset*. Naturalmente, uma palavra pode ter diversos sinônimos a depender do contexto. Assim, o Tep 2.0 traz os contextos expressos por meio de um número sequencial.

- *Wordlist* do Corpop (PASQUALINI, 2018): uma lista de palavras com seus respectivos números de frequência, obtidos por um compilado de textos do PB popular escrito e selecionados com base no nível de letramento médio do país.

- PortiLexicon (USP, S. C., 2022): contém mais de 1,2 milhão de formas de palavras em português com suas respectivas informações morfológicas e morfossintáticas, seguindo o modelo internacional de Dependências Universais. O léxico é baseado no Unitex-PB(MUNIZ, 2004) e faz parte do projeto POeTiSA(USP - SÃO CARLOS, 2022).

Estes conjuntos foram combinados conforme demonstrado na Figura 2, de modo que foi gerado um léxico complementar, ou Contexto de Palavras, composto por 12.564 palavras, dentre verbos, adjetivos, advérbios e substantivos, com informações de: (a) sinônimo mais simples, (b) forma no infinitivo, (c) gênero e (d) número.

A metodologia para construção deste léxico de itens e sinônimos considerou a combinação de cada palavra e seu conjunto de sinônimos apresentados no TeP 2.0, em suas formas no infinitivo (advindas da base léxica - PortiLexicon), com a frequência de cada palavra correspondente no Corpop, também em sua forma no infinitivo dada pelo PortiLexicon. Assim, o sinônimo com maior número de *hits* do Corpop, *i.e.* maior frequência, foi eleito como sinônimo mais simples. Na sequência, foram incluídas informações de gênero e número das palavras a partir do PortiLexicon para facilitar a SS, garantindo coesão e coerência textual. As substituições automaticamente realizadas nesta metodologia consideraram a substituição de uma palavra flexionada em um determinado gênero e grau por

outra na mesma flexão. Em casos de impossibilidade sintática, ou seja, nos casos em que não há um sinônimo correspondente para a mesma flexão de gênero e número, optou-se pela não substituição. Ainda que este comportamento restrinja o número de substituições possíveis, a redução na complexidade do processo de substituição é significativa. Isto porque uma mudança de flexão em uma única palavra pode implicar na alteração de flexão em diversas palavras dependentes em diferentes sentenças do texto.

Para facilitar o entendimento da construção deste léxico, podemos dizer que a solução foi embasada em três macro etapas:

1. Adicionar informação de genêro/número vindas do Portilexicon (USP, S. C., 2022) na lista de palavras do CorPop (PASQUALINI, 2018) e do Tep 2.0 (DIAS-DA-SILVA; MORAES, 2003);

2. Consolidar os dois conjuntos enriquecidos para captar número de *hits* cada palavra do Tep 2.0;

3. Identificar para dado gênero e número, a palavra sinônima com maior número de *hits*.

Consideremos o exemplo do lema PROFESSOR. No Corpop, a relação das palavras que contém este lema está representada na Figura 3.

```
df_corpop[df_corpop['words'].str.contains('professor')]
```

	hits	words
331	250	professor
1702	43	professores
2409	29	professora
13727	3	professoras
29309	1	professore

Figura 3: Relação de palavras com lema “professor” no Corpop

Observamos que a palavra PROFESSOR apresenta maior número de *hits* na lista de palavras do Corpop. Isto significa que é mais comum no vocabulário popular brasileiro do que as demais flexões do lema PROFESSOR. É possível observar também que, possivelmente devido a um erro de digitação, um dos textos do Corpop contém a palavra PROFESSORE com apenas uma ocorrência no *corpus*. Como se verifica na Figura 4, devido ao fato desta palavra ser inexistente no Portilexicon, permanecem apenas as palavras

professor, professores, professora e professoras com suas respectivas indicações de gênero e número. Como estas palavras contêm o mesmo lema [4], foi mantido (como indicado na Figura 4) o maior número de *hits* dado pela palavra PROFESSOR na Figura 3.

```
df_corpop_lexicon[df_corpop_lexicon['infinitive']=='professor']
        words       infinitive gender number hits
155648  professor   professor  Masc   Sing   250.0
155649  professora  professor  Fem    Sing   250.0
155658  professoras professor  Fem    Plur   250.0
155659  professores professor  Masc   Plur   250.0
```

Figura 4: Combinação das informações de gênero e número do Portilexicon com o Corpop para o lema "professor"

O mesmo processo que foi realizado para o Corpop, foi feito para o Tep 2.0. Como mencionado anteriormente, o Tep 2.0 é baseado em contextos. Então, consideremos o exemplo do contexto dado pela identificação da chave 18796, representado na Figura 5.

```
df_tep[df_tep['key'] == 18796]
       key   type          infinitive
70622  18796  Substantivo  mestre
70623  18796  Substantivo  prelecionador
70624  18796  Substantivo  preletor
70625  18796  Substantivo  professor
```

Figura 5: Relação de palavras para a chave de contexto 18796 no Tep

Neste determinado contexto, verifica-se que a palavra PROFESSOR pode ser substituída por MESTRE, PRELECIONADOR ou PRELETOR. O próximo passo é combinar as formas no infinitivo dadas pelo Tep 2.0 com as flexões proporcionadas pelo Portilexicon, o resultado é visto na Figura 6. Assim, temos a forma no infinitivo com suas possíveis flexões de gênero e número.

A próxima macro etapa é a de consolidar todas as informações presentes nos resultados das combinações anteriores, vide Figura 7. Desta forma, conseguimos visualizar para um contexto específico do Tep quais são as palavras relacionadas, com suas formas flexionadas e no infinitivo, bem como o número de ocorrências no Corpop.

Por fim, para cada palavra presente na relação, a depender do seu gênero e número, é identificado qual o sinônimo do contexto com maior número de ocorrências (ou *hits*) do

[4] Os lemas das palavras estão discriminados pela coluna de nome *infinitive*.

```
df_tep_lexicon[df_tep_lexicon['key']==18796]
       key   type        infinitive    words           gender number
158598 18796 Substantivo mestre        mestra          Fem    Sing
158600 18796 Substantivo mestre        mestras         Fem    Plur
158602 18796 Substantivo mestre        mestre          Masc   Sing
158604 18796 Substantivo mestre        mestres         Masc   Plur
158606 18796 Substantivo prelecionador prelecionador   Masc   Sing
158607 18796 Substantivo prelecionador prelecionadores Masc   Plur
158609 18796 Substantivo professor     professor       Masc   Sing
158610 18796 Substantivo professor     professora      Fem    Sing
158611 18796 Substantivo professor     professoras     Fem    Plur
158612 18796 Substantivo professor     professores     Masc   Plur
```

Figura 6: Combinação das informações de gênero e número do Portilexicon com o Tep para a chave de contexto 18796

```
df_tep_hits[df_tep_hits['key']==18796]
       key   type        infinitive    words           gender number hits
1285   18796 Substantivo professor     professora      Fem    Sing   250.0
1286   18796 Substantivo professor     professores     Masc   Plur   250.0
1287   18796 Substantivo professor     professor       Masc   Sing   250.0
1288   18796 Substantivo professor     professoras     Fem    Plur   250.0
4109   18796 Substantivo mestre        mestres         Masc   Plur   94.0
4110   18796 Substantivo mestre        mestra          Fem    Sing   94.0
4111   18796 Substantivo mestre        mestras         Fem    Plur   94.0
4112   18796 Substantivo mestre        mestre          Masc   Sing   94.0
107526 18796 Substantivo prelecionador prelecionador   Masc   Sing   NaN
107527 18796 Substantivo prelecionador prelecionadores Masc   Plur   NaN
```

Figura 7: Consolidação das combinações com Corpop, Tep e Portilexicon

Corpop. Além disto, são removidos os casos em que tanto o sinônimo quanto a palavra são idênticos. Assim, no caso de uma SS de um texto simples, caso apareça a palavra MESTRA, esta poderá ser substituída por PROFESSORA, ou alguma outra variação, caso a palavra MESTRA esteja presente em outra chave de contexto. No caso de uma SS de um texto complexo, o processo é invertido: caso apareça a palavra PROFESSORA esta poderá ser substituída por MESTRA, ou alguma outra variação a depender dos contextos envolvendo a palavra PROFESSORA.

Observa-se, portanto que o Corpop apresenta um papel de extrema importância para o formato de SS aqui apresentado, haja vista que:

- Foi desenvolvido a partir da análise de dados sobre o nível de letramento dos leitores brasileiros e das características que poderiam compor um padrão de simplicidade

```
df_context[df_context['key']==18796]
```

	key	type	infinitive	words	gender	number	hits	replace	hits_replace
398	18796	Substantivo	mestre	mestre	Masc	Sing	94.0	professor	250.0
399	18796	Substantivo	mestre	mestres	Masc	Plur	94.0	professores	250.0
400	18796	Substantivo	mestre	mestra	Fem	Sing	94.0	professora	250.0
401	18796	Substantivo	mestre	mestras	Fem	Plur	94.0	professoras	250.0
48212	18796	Substantivo	prelecionador	prelecionador	Masc	Sing	NaN	professor	250.0
48213	18796	Substantivo	prelecionador	prelecionadores	Masc	Plur	NaN	professores	250.0

Figura 8: Contexto de palavras filtrado para a chave de contexto 18796

```
df_context[df_context['key']==18796]
```

	key	type	infinitive	words	gender	number	hits	replace
750	18796	Substantivo	professor	professoras	Fem	Plur	250.0	professores
751	18796	Substantivo	professor	professor	Masc	Sing	250.0	professores
752	18796	Substantivo	professor	professora	Fem	Sing	250.0	professores
2640	18796	Substantivo	mestre	mestras	Fem	Plur	94.0	professores
2641	18796	Substantivo	mestre	mestra	Fem	Sing	94.0	professores
2660	18796	Substantivo	mestre	mestres	Masc	Plur	94.0	professores
2664	18796	Substantivo	mestre	mestre	Masc	Sing	94.0	professores
75379	18796	Substantivo	prelecionador	prelecionador	Masc	Sing	NaN	professores
75380	18796	Substantivo	prelecionador	prelecionadores	Masc	Plur	NaN	professores

Figura 9: Contexto de palavras independente de flexão filtrado para a chave de contexto 18796

textual em um *corpus* adequado a estes leitores (PASQUALINI, 2018), estando diretamente relacionado aos objetivos motivadores desta pesquisa;

- Apresenta dados de frequência para cada uma das palavras consideradas mais simples ou comuns no vocabulário popular brasileiro. Conforme um dos principais pontos de análise apresentados por Wilkens, Vecchia et al. (2014), ao contrário do que geralmente se supõe, o tamanho de uma palavra não é um bom indicativo para categorização de palavras simples e complexas, mas sim, a frequência de sua utilização.

Assim sendo, a partir do léxico criado, é feita a varredura de cada um dos textos contidos no *corpus* de treinamento (vide Figura 2), em que são verificadas todas as palavras dos textos por meio do *tokenizador* da biblioteca NLTK(BIRD; KLEIN; LOPER, 2009). Apenas os *tokens* identificados no *corpus* e que não fossem nomes próprios dados pela biblioteca Spacy (HONNIBAL; MONTANI, 2017), são passíveis de serem substituídos por seus respectivos sinônimos.

Devido ao fato de existirem diferentes contextos de sinônimos expressos pelo Tep 2.0,

uma mesma palavra pode ser substituída por mais de um sinônimo. Nesse sentido, a metodologia proposta considerou uma seleção inspirada no algoritmo da roleta (SHUKLA; PANDEY; MEHROTRA, 2015). É importante ressaltar que por se tratar de uma estratégia usada para AD, é essencial que a SS desenvolvida seja capaz de manter o rótulo inicial do texto, *i.e.* simples ou complexo. Portanto, no caso de um texto originalmente simples, os sinônimos com maior frequência no CorPop possuíam uma porção maior da roleta, e aqueles com frequência mais baixa possuíam uma porção relativamente menor. O processo inverso válido foi realizado para os textos originalmente complexos.

Uma vez definida a proporção dos itens da roleta, o algoritmo roda um determinado número de vezes (conforme parâmetro informado pelo usuário; neste trabalho foi definido o número 5), de modo que a cada iteração é escolhido apenas um sinônimo para cada *token*. Considerando a analogia, a roleta é girada um determinado número de vezes, e os sinônimos escolhidos são aqueles sorteados na roleta. Para o sorteio, cada sinônimo i do conjunto de contextos recebe um percentual de probabilidade dado por: $P(i) = \frac{\texttt{número de hits}_i}{\sum \texttt{número de hits}}$. Durante o sorteio, um dos sinônimos é amostrado aleatoriamente; caso a probabilidade $P(i)$ do sinônimo seja inferior a um número também gerado aleatoriamente, o sinônimo é eleito para substituição. Este processo ocorre de forma indefinida até que haja um eleito. Consideremos o texto *Escavação Arqueológica* indicado abaixo:

> *Escavação Arqueológica*
> *A escavação arqueológica é a forma usada pelos arqueólogos para encontrar as coisas do passado e, a partir delas, contar histórias sobre quem as fabricou, usou e descartou.*
> *O local onde essas coisas são encontradas – chamado de sítio arqueológico, é dividido em quadrados ("quadrículas") de tamanho variável, geralmente de um metro quadrado. É nas quadrículas que os arqueólogos escavam. Esse processo é realizado por níveis (quando a escavação é organizada em centímetros) ou por camadas (quando a escavação é organizada por tipos diferentes de solos).*
> *Quando um objeto é encontrado, escava-se cuidadosamente a área ao seu redor, fotografando-o, descrevendo-o e coletando-o, preservando as informações de sua exata localização.*
> *Este espaço é uma representação de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).*

Por se tratar de um texto classificado como complexo, a orientação da SS será por

substituir as palavras por aquelas com menor número de ocorrências no Corpop, de modo a minimizar possibilidades de alteração do rótulo original do texto.

O primeiro passo da substituição é converter o texto em uma lista de sentenças, vide Figura 10. Esta lista é iterada, de modo que cada sentença é *tokenizada* por meio da biblioteca NLTK, e cada *token* é avaliado por meio da biblioteca Spacy para identificar se constitui um nome próprio ou não. A ideia em identificar nomes próprios surgiu após a primeira tentativa de substituição, em que foi verificado no texto *A Viagem do Beagle*, a substituição da localidade *Cabo Verde* por *rabo verde*. Por se tratar de uma alteração significativa no contexto do texto, optou-se por inviabilizar as substituições no caso de nomes próprios.

```
sentences
['Escavação Arqueológica ', 'A escavação arqueoló...descartou.', 'O local onde essas
c...de solos).', 'Quando um objeto é e...calização.', 'Este espaço é uma re...alma (R
S).']
> special variables
> function variables
0: 'Escavação Arqueológica'
1: 'A escavação arqueológica é a forma usada pelos arqueólogos para encontrar as coisas
do passado e, a partir delas, contar história...'
2: 'O local onde essas coisas são encontradas chamado de sítio arqueológico, é dividido
em quadrados ("quadrículas") de tamanho va...'
3: 'Quando um objeto é encontrado, escava-se cuidadosamente a área ao seu redor, fotogra
fando-o, descrevendo-o e coletando-o, preser...'
4: 'Este espaço é uma representação de escavação arqueológica no Abrigo de Canhemborá, s
ituado no município de Nova Palma (RS).'
len(): 5
```

Figura 10: Lista de sentenças

```
word_tokenize(sentence)
> ['Este', 'espaço', 'é', 'uma', 'representação', 'de', 'escavação', 'arqueológica', 'n
o', 'Abrigo', 'de', 'Canhemborá', ',', 'situado', ...]
ner
> ['Abrigo', 'Canhemborá', 'Nova', 'Palma', 'RS']
```

Figura 11: Lista de *tokens* e nomes próprios

Consideremos a última sentença do exemplo que pode ser verificada na Figura 10, "*Este espaço é uma representação de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Plama (RS)*". Os *tokens* indicados pela lista denominada *ner* da imagem 11, foram identificados pela biblioteca Spacy como nomes próprios, assim, não são passíveis de substituição. Dentro do processo, cada *token* da sentença consulta

o léxico dado pelo contexto de palavras. Tomemos como exemplo a segunda palavra da última sentença, *i.e.*, *espaço*. As possibilidades de substituição, filtradas por gênero e número de acordo com a flexão da palavra *espaço*, dentro do contexto de palavras estão indicadas na Figura 12.

df_selection

	key	type	infinitive	words	gender	number	hits_replace	replace	hits
1192	13606	Substantivo	aberto	aberto	Masc	Sing	42.0	ESPAÇO	107.0
1936	18548	Substantivo	tamanho	tamanho	Masc	Sing	27.0	ESPAÇO	107.0
2239	13606	Substantivo	afastamento	afastamento	Masc	Sing	23.0	ESPAÇO	107.0
2980	15620	Substantivo	trecho	trecho	Masc	Sing	18.0	ESPAÇO	107.0
3055	11759	Substantivo	termo	termo	Masc	Sing	18.0	ESPAÇO	107.0
7763	11531	Substantivo	intervalo	intervalo	Masc	Sing	5.0	ESPAÇO	107.0
8143	15620	Substantivo	intervalo	intervalo	Masc	Sing	5.0	ESPAÇO	107.0

Figura 12: Contexto de palavras considerando a substituição da palavra *espaço*

Com base nesta seleção, é calculada a probabilidade para cada uma das palavras substitutas possíveis, vide Figura 13. Para escolha da palavra substituta é utilizado o algoritmo da roleta, de modo que seguindo a parametrização inicial definida pelo usuário, são criados cinco textos a partir do texto original. Isto significa que a "roleta" é "girada" cinco vezes para cada um dos *tokens* passíveis de substituição do texto original e cada substituto eleito compõe uma das cinco saídas possíveis.

Por se tratar de um texto complexo, as palavras com probabilidade mais baixa são preferidas na aleatoriedade amostral seguindo a lógica dos algoritmos genéticos (SHUKLA; PANDEY; MEHROTRA, 2015). Conforme verificado na Figura 13, as palavras eleitas para substituição apresentam uma menor probabilidade de ocorrência pelo Corpop. Ao final do processo, temos cinco variações da sentença:

Tabela 6: Variações da última sentença do texto *Escavação Arqueológica* com classe complexa por SS

Original	**Este espaço** é uma **representação** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).
Versão 1	**Nascente intervalo** é uma **simulação** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).

Versão 2	**Nascente tamanho** é uma **reprodução** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).
Versão 3	**Nascente termo** é uma **reprodução** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).
Versão 4	**Nascente tamanho** é uma **exibição** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).
Versão 5	**Nascente trecho** é uma **simulação** de escavação arqueológica no Abrigo de Canhemborá, situado no município de Nova Palma (RS).

4.2 Retrotradução (RT)

Conforme Ferreira e Costa (2020), a primeira vez que o método RT foi utilizado como uma solução para AD, foi no trabalho de Yu et al. (2018). Esse método visa gerar, automaticamente, sentenças com o mesmo significado e classificação das sentenças originais por meio da tradução de uma língua X para uma língua Y, com uma retrotradução para a língua X.

Conforme demonstrado por Ferreira e Costa (2020), o método tradicional de RT normalmente gera sentenças com estrutura gramatical correta e que tendem a capturar de forma mais realista o verdadeiro significado das sentenças. Neste sentido, haja vista que a leiturabilidade de um texto está fortemente atrelada à estrutura semântica e sintática das sentenças (TAYLOR, 1953), optou-se por avaliar os impactos da inclusão de textos gerados artificialmente pelo método tradicional de RT. Para isto, foi utilizada a biblioteca *Deep Translator* (BACCOURI, 2020), que é uma ferramenta gratuita compatível com Python integrada a diversos tradutores. Neste trabalho, foi utilizada a integração com o Google Tradutor (GOOGLE, 2023).

O processo em si de construção da tarefa é bem simples, haja vista que a biblioteca *Deep Translator* permite a parametrização do idioma de origem para o idioma desejado e não há restrição no número de palavras.

No entanto, há uma preocupação na escolha do idioma intermediário, *i.e.* da língua Y,

```
df_selection
    replace     hits  probability
0   ABERTO      42.0  0.315789
44  AFASTAMENTO 23.0  0.172932
62  INTERVALO   5.0   0.037594
63  TAMANHO     27.0  0.203008
70  TERMO       18.0  0.135338
75  TRECHO      18.0  0.135338

replacement_word
'intervalo'

replacement_word
'tamanho'

replacement_word
'termo'

replacement_word
'tamanho'

replacement_word
'trecho'
```

Figura 13: Cálculo da probabilidade das palavras que podem substituir a palavra *espaço*, seguido das palavras eleitas a cada iteração da roleta

porque afeta diretamente o resultado do processo de RT e, consequentemente, a manutenção do rótulo inicial do texto. Ao mesmo tempo, determinar se um idioma é mais simples ou complexo que outro é uma questão controversa e dependente do contexto; inclusive, estudos recentes apontam que não é possível medir se uma língua é mais complexa que outra (BENTZ et al., 2022).

Neste sentido, para a escolha dos idiomas intermediários envolvidos na tarefa, foi considerado o fato de que as línguas de família românica, como o francês, espanhol, italiano e o português, tendem a ter um espectro maior de variáveis no plano sintático, em termos de gênero gramatical, declinação de substantivos e adjetivos, bem como conjugações verbais, enquanto as línguas anglosaxônicas, como o inglês, o alemão e o holandês, tendem a ser mais simples nesse sentido (PEI; GAYNOR, 1954). Por sua vez, o italiano pode ser considerada a língua que conserva mais palavras com maior proximidade das palavras latinas, logo com um maior número de declinações; e o inglês, além de ser um idioma com elevado número de recursos computacionais, é conhecida por ter poucas modificações na conjugação de verbos e inexistência de gêneros gramaticais.

Nesse contexto de especificidades idiomáticas e gramaticais envolvidas, realizou-se a varredura nos textos do *corpus* de treinamento, considerando seus respectivos rótulos

originais para identificação da língua intermediária. Assim, optou-se pela RT com a língua inglesa como intermediária dos textos simples, e da língua italiana para os textos complexos.

Para que possamos compreender como variam os textos aumentados para a classe simples e aqueles para a classe complexa, consideremos os casos dos textos *Batatas Energéticas* e *Alma de Gato* apresentados na Tabela 7.

Tabela 7: Comparativo entre textos originais e aumentados por RT

Rótulo	**Texto Original**	**RT**
Simples	BATATAS ENERGÉTICAS **Aperte o botão** e observe o relógio. **Ele** começou a funcionar por causa das batatas. Por que isso acontece? Em cada batata há uma **barrinha** de zinco e outra de cobre. As reações químicas dentro das batatas fazem com que haja uma diferença de energia entre as **barrinhas**. Essa diferença produz eletricidade suficiente para fazer **o relógio funcionar**. E o que isso tem a ver com a sua vida? Esse fenômeno também acontece quando você usa um **celular** ou outro **aparelho** que funcione com **pilhas ou baterias**. Dentro da pilha, **acontece** uma reação química **bem** parecida **com a que acontece** nas batatas. Essa reação produz uma tensão elétrica entre suas **pontas**.	BATATAS ENERGÉTICAS **Pressione** o botão e observe o relógio. Começou a funcionar por causa das batatas. Por que isso acontece? Em cada batata há uma **barra** de zinco e outra de cobre. As reações químicas dentro das batatas fazem com que haja uma diferença de energia entre as **barras**. Essa diferença produz eletricidade suficiente para fazer **funcionar o relógio**. E o que isso tem a ver com a sua vida? Esse fenômeno também acontece quando você usa um **telefone celular** ou outro **dispositivo** que funcione **com baterias**. Dentro da pilha, **ocorre** uma reação química **muito** parecida **com a que ocorre** nas batatas. Essa reação produz uma tensão elétrica entre suas **extremidades**.

Tabela 7: Comparativo entre textos originais e aumentados por RT (Continuação)

Rótulo	**Texto Original**	**RT**
Complexo	BATATAS ENERGÉTICAS **Aperte o botão** e observe o relógio. As batatas **colocaram-no em funcionamento**. Por que isso acontece? Observe que **em cada batata foram fixados um** terminal de zinco e **outro** de cobre. Reações químicas que acontecem dentro das batatas provocam uma diferença de potencial entre os terminais, **o que produz energia elétrica** suficiente para fazer o relógio funcionar. E o que isso tem a ver com a sua vida? Se você já utilizou um **celular** ou qualquer outro **aparelho alimentado por** pilha ou bateria, já fez uso deste fenômeno. **Da mesma maneira que ocorre nas** batatas, uma reação química planejada **acontece** dentro da pilha, produzindo uma diferença de potencial entre suas **extremidades**.	BATATAS ENERGÉTICAS **Pressione o botão** e observe o relógio. As batatas **funcionaram**. Por que isso acontece? Observe que **cada batata tem** um terminal de zinco e **um** de cobre conectados. **As** reações químicas que ocorrem dentro das batatas provocam uma diferença de potencial entre os terminais, **que produz eletricidade** suficiente para fazer o relógio funcionar. E o que isso tem a ver com a sua vida? Se você já usou um **telefone celular** ou qualquer outro **dispositivo movido a** pilha ou bateria, já fez uso desse fenômeno. **Assim como acontece com as** batatas, uma reação química planejada **ocorre** dentro da pilha, produzindo uma diferença de potencial entre suas **pontas**.

Tabela 7: Comparativo entre textos originais e aumentados por RT (Continuação)

Rótulo	**Texto Original**	**RT**
Simples	Alma-de-gato (Piaya cayana) **Locais** onde **vive** **Ramos** de árvores em matas, capoeiras, parques e, **inclusive**, nas cidades. Características Tem cor **castanha** e olhos vermelhos. Também tem uma cauda longa. Muitas pessoas veem essa ave pulando entre os galhos das árvores **atrás de** comida. **Por conta** dessas características, são **parecidas com** um esquilo. Os machos e as fêmeas **dessa** espécie são **bem parecidos**. **De que** se alimenta? Insetos, principalmente lagartas, inclusive as venenosas. Pode comer pequenos vertebrados **também**, como lagartixas, pererecas, **aves** e ovos. Como **é** seu ninho? Os galhos que formam os ninhos são **bem** frágeis. Mas os ninhos podem conter até seis ovos. Tanto os machos quanto as fêmeas cuidam dos filhotes. Como é seu canto? Ouvimos o seu canto principalmente no início da primavera, quando começa o período de reprodução.	Alma de Gato (Piaya cayana) **Lugares** onde **você mora** **Galhos** de árvores em matas, capoeiras, parques e **até mesmo** nas cidades. Características Tem cor **marrom** e olhos vermelhos. **Ele** também tem uma longa cauda. Muitas pessoas veem essa ave pulando entre os galhos das árvores **em busca** de comida. **Por causa** dessas características, **ele**s são **semelhantes a** um esquilo. Os machos e as fêmeas **desta** espécie são **muito semelhantes**. **Do que ele** se alimenta? Insetos, principalmente lagartas, inclusive venenosas. **Também** pode comer pequenos vertebrados, como lagartixas, pererecas, **pássaros** e ovos. Como **está** o seu ninho? Os galhos que formam os ninhos são **muito** frágeis. Mas os ninhos podem conter até seis ovos. Tanto os machos quanto as fêmeas cuidam dos filhotes. Como é o seu canto? Ouvimos seu canto principalmente no início da primavera, quando começa o período de reprodução.

Tabela 7: Comparativo entre textos originais e aumentados por RT (Continuação)

Rótulo	**Texto Original**	**RT**
	É **bem** parecido com o gemido de um gato. Por isso recebeu o nome popular de alma-de-gato.	É **muito** parecido com o gemido de um gato. Por isso recebeu o nome popular de alma de gato.
Complexo	Alma-de-gato (Piaya cayana) Onde pode ser **vista?** Pode ser facilmente observada **nos ramos** de árvores em **matas**, capoeiras, parques e, **inclusive**, nas cidades. Como pode ser **reconhecida**? Essa ave, de **coloração castanha** e olhos vermelhos, **muitas vezes é observada** pulando entre os galhos das árvores **atrás** de alimento. Esse comportamento, somado à longa cauda escalonada que possui, assemelha a espécie a um esquilo, razão **de seu nome em inglês**. **Os** machos e as fêmeas dessa espécie da família **dos cuculídeos** são **semelhantes fisicamente**. De que se alimenta? Insetos, principalmente lagartas, inclusive **as** venenosas. Pode comer pequenos vertebrados **também**, como lagartixas, pererecas, **aves** e ovos. Como **é** seu ninho? É uma estrutura **de galhos frouxa**, que pode conter até seis ovos.	Alma de gato (Piaya cayana) Onde pode ser **visto**? Pode ser facilmente observada **em galhos** de árvores em **florestas**, capoeira, parques e **até** cidades. Como pode ser **reconhecido**? Essa ave, de **cor marrom** e olhos vermelhos, **costuma ser vista** pulando entre os galhos das árvores **em busca** de alimento. Esse comportamento, somado à longa cauda escalonada que possui, assemelha a espécie a um esquilo, razão **pela qual em inglês recebe esse nome**. Machos e fêmeas desta espécie da família **Cuculidae** são **fisicamente semelhantes**. Do que **ele** se alimenta? Insetos, principalmente lagartas, inclusive venenosas. **Também** pode comer pequenos vertebrados, como lagartixas, pererecas, **pássaros** e ovos. Como **está** o seu ninho? É uma estrutura **ramificada solta**, que pode conter até seis ovos.

Tabela 7: Comparativo entre textos originais e aumentados por RT (Continuação)

Rótulo	**Texto Original**	**RT**
	Tanto os machos quanto as fêmeas cuidam dos filhotes. Como é seu canto? **É ouvido** principalmente no início do período reprodutivo – **a** primavera – e **assemelha-se** ao gemido de um gato, por isso recebeu o nome popular de alma-de-gato.	Tanto os machos quanto as fêmeas cuidam dos filhotes. Como é **o** seu canto? **Ouve-se** principalmente no início do período reprodutivo - primavera - e **lembra o** gemido de um gato, por isso recebeu o nome popular de alma de gato.

Identifica-se que, apesar da possibilidade de inclusão de ruídos (como quando substitui a sentença "Locais onde vive" por "Lugares onde você mora" em um texto sobre a espécie Alma de Gato), o processo de RT acaba por estruturar de forma mais padronizada a sintaxe das sentenças e considera expressões e/ou palavras aparentemente mais comuns no vocabulário do PB. Assim, de forma simplista, o processo de RT se aproxima de uma ST, alguns exemplos que demonstram esta argumentação são:

- o fato de que independente da língua intermediária, a escolha do vocabulário é muito próxima para as classes simples/complexa. Isto fica evidente no trecho que se lê sobre a cor *castanha* que foi sempre substituída por *marrom* após a RT. Assim como, *aperte* por *pressione*, *inclusive* por *até*, *atrás* por *em busca*.
- após a RT as sentenças passam a seguir um formato mais estruturado em termos sintáticos, com a inclusão de artigos e/ou pronomes, por exemplo, *Reações químicas que acontecem dentro das batatas* virou *As reações químicas que ocorrem dentro das batatas*, *Também tem uma cauda longa* virou *Ele também tem uma cauda longa*, *Como é seu canto?* virou *Como é o seu canto?*.

4.3 Análise holística dos resultados

De forma geral, observa-se que o método de SS acaba por substituir uma quantidade restrita de palavras, cerca de 19% das palavras dos textos do *corpus* original do Museu. A média de substituição para os textos difíceis é de quase 25%, e nos textos simples

este valor cai para 13%. Considerando o par do texto *Escavação Arqueológica* na versão simples, apresentado abaixo:

> *Escavação Arqueológica*
> *Os arqueólogos, profissionais que investigam e analisam objetos do passado, utilizam a escavação arqueológica para fazer seu trabalho. Com a escavação, eles encontram as coisas do passado e podem contar histórias sobre quem as fabricou, usou e jogou fora.*
> *Eles encontram essas coisas num lugar chamado sítio arqueológico, que é dividido em quadrados ("quadrículas"). Os tamanhos variam, mas geralmente são de um metro quadrado. É nas quadrículas que os arqueólogos escavam. Eles organizam as escavações de diversas maneiras. Pode ser por níveis (quando a escavação é organizada em centímetros) ou por camadas (quando a escavação é organizada por tipos diferentes de solos).*
> *Quando encontram um objeto, eles escavam com muito cuidado a área em volta. Também tiram fotos, descrevem e coletam os materiais para preservar as informações de sua exata localização.*
> *Este espaço é uma representação de escavação arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.*

as variações para a última sentença foram:

Tabela 8: Variações da última sentença do texto *Escavação Arqueológica* com classe simples por SS

Original	Este **espaço** é uma **representação** de **escavação** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.
Versão 1	Este **lugar** é uma **imagem** de **cova** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.
Versão 2	Este **lugar** é uma **imagem** de **cova** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.
Versão 3	Este **lugar** é uma **expressão** de **perfuração** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.
Versão 4	Este **lugar** é uma **expressão** de **cova** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.

Versão 5	Este **lugar** é uma **expressão** de **cova** arqueológica no Abrigo de Canhemborá, em Nova Palma, no Rio Grande do Sul.

Nota-se que no caso da SS para textos simples, o número de possibilidades de palavras se torna ainda mais restrito. Enquanto na Tabela 6 o espectro das substituições alcançou 8 palavras diferentes, no caso dos textos simples representados na Tabela 8 este número caiu para 5. Isto está relacionado ao fato de que a palavra a ser substituída não pode ser mais complexa do que a palavra escolhida pelo autor, que por sua vez atuou com foco na simplificação do texto.

Também é importante destacar que os textos aumentados por RT apesar de nem sempre apresentarem a melhor escolha de palavras ou expressões, apresentavam maior nível de manutenção da coesão e coerência textuais do que aqueles aumentados por SS. Tomando como exemplo o texto *Alma-de-gato*, um dos casos de SS para a classe complexa gerou o seguinte trecho:

Onde pode **haver inspeção**?
Pode **haver facil** observada nos **ramais** de árvores em matas, capoeiras, parques e, **inclusiva**, nas **urbes**.
Quão pode **nascer consagrada**?
Essa ave, de **pigmentação** castanha e **orifícios incendidos**, **demasiadas margens** é observada pulando entre os galhos das árvores **anterior** de **alento**.

E para a classe simples:

Locais onde vive
Ramos de árvores em **florestas**, capoeiras, parques e, inclusive, nas cidades.
Próprias
Tem cor castanha e olhos **encarnados**. **Mais** tem uma cauda **longamente**. Muitas pessoas veem essa ave pulando entre os **ramos** das árvores **depois** de comida. Por conta dessas **próprias**, são **mesmas** com um esquilo. Os **homens** e as **mulheres** dessa espécie são bem **mesmos**.

Observa-se que, embora as alterações tenham sido realizadas de forma bem pontual, dificultou-se enormemente a compreensão do sentido da mensagem uma vez que a escolha dos sinônimos não considera o contexto de substituição.

5 Metodologia Experimental

A abordagem proposta neste trabalho desenvolve uma análise embutida de mecanismos de AD em termos de leiturabilidade para o PB que foram apresentados no Capítulo 4. Os métodos de AD desenvolvidos geraram dois *corpora* artificiais a partir da extração de dois corpora criados manualmente, compreendendo um total de 504 textos sintéticos conforme pode ser observado na Tabela 9.

Tabela 9: *Corpora* considerados na metodologia proposta

		Classe Simples		**Classe Complexa**	
		Número de textos	**Média de palavras por texto**	**Número de textos**	**Média de palavras por texto**
Corpus principal	**Museu**	42	164	42	158
Corpus aumentado	**SS**	210	164	210	158
	RT	42	161	42	155
Corpus secundário	**Wikibooks**	15	1620	62	1764

A fim de avaliar os resultados deste aumento, definem-se quatro etapas principais representadas na Figura 14, explicadas em mais detalhes na sequência. Para facilitar o entendimento do processo como um todo, foi incluída a etapa de aplicação dos métodos de AD delineada no Capítulo anterior na Figura 14.

Para o desenvolvimento da análise proposta, utilizou-se a linguagem *Python*, de modo que cada uma das etapas consistiu em múltiplas tarefas com uso de diferentes bibliotecas de código aberto. Em termos de arquitetura, para facilitar a identificação na armazenagem das informações contidas, seguiu-se a estrutura de *medallion* (LEE; HEINTZ, 2019), que consiste no armazenamento em camadas:

- *Raw*: contém dados brutos, isto é, os textos de entrada dos *corpora* para as etapas de coleta e enriquecimento;
- *Bronze*: contém os dados brutos estruturados em único arquivo, agregados aos textos gerados artificialmente pós etapa de AD;

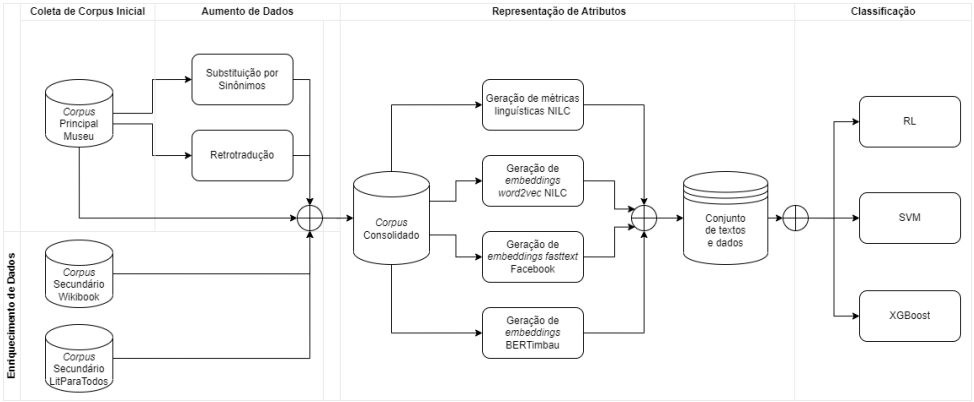


Figura 14: Metodologia proposta para avaliação dos impactos da inclusão de dados sintéticos em classificadores automáticos de leiturabilidade textual

- *Silver*: contém os textos da camada bronze consolidados com suas respectivas representações de atributos;
- *Gold*: contém os dados agregados da camada *silver*, normalizados para entrada do modelo de classificação. A normalização do conjunto de dados consiste em um passo importante no pré-processamento para a etapa de agrupamento. Esse é um requisito comum para muitos modelos de AM. Isto porque, quando uma característica, entre várias, possui ordem de magnitude maior que outras, ela pode dominar a função objetivo e tornar o modelo incapaz de aprender com outras características conforme esperado. Desta forma, a padronização foi realizada por meio do pacote Scikit-Learn (SCIKIT-LEARN, 2021b), que transforma a média amostral para zero e dimensiona a variância da amostra para unitária.

Os dados gerados para cada camada foram salvos em **.parquet*, um formato de armazenamento de dados gratuito e de código aberto orientado a colunas, que permite uma redução de espaço de armazenamento, execução mais rápida e fácil integração com a linguagem Python (VANROSSUM; DRAKE, 2010).

5.1 Coleta de *Corpora*

Esta seção apresenta os *corpora* considerados na metodologia ilustrados na Figura 14. O processo de coleta dos textos dos *corpora* selecionados atende à camada *raw* da estrutura de *medallion*. De modo que, os textos brutos foram extraídos de fontes com extensões diferentes (**.txt*, **.tsv* e **.vert*) e inseridos em um *DataFrame* (PANDAS, 2022) com a seguinte estrutura:

key	*Title*	*Corpus*	*Label*	*Content*

Dado que:

- A coluna *key* representa uma chave de identificação única para cada um dos textos. Ela foi criada como um recurso de desenvolvimento com o intuito de facilitar a mesclagem dos dados para as etapas posteriores;
- A coluna *Title* apresenta o título do texto dado pelo nome do arquivo;
- A coluna *Corpus* identifica o nome do *corpus* de origem;
- A coluna *Label* indica o rótulo/classe que o texto pertence, *i.e.* simples ou complexo. Destaca-se que os rótulos foram descritos inicialmente no título do arquivo, caso o título contivesse a palavra `original` tratava-se da versão inicial do texto, logo associado ao rótulo `complexo`. No caso dos textos cujo rótulo seria `simples`, os títulos continham a palavra `simplificado`;
- A coluna *Content* contém os textos em si com os devidos tratamentos envolvendo a eliminação de componentes não-textuais, como números de páginas, identificação de seções e hiperlinks.

5.1.1 *Corpus* principal

Haja vista que não há um consenso entre os estudiosos da área de linguagem sobre as métricas de análise linguística existentes em termos de categorização textual por leiturabilidade (SANTOS, 2010; SCARTON; ALUÍSIO, 2010; FINATTO, 2020) - uma vez que estas não se relacionam às propriedades conceituais do texto e não consideram a interação entre texto e leitor; é essencial que a escolha e classificação do conjunto inicial de textos para treino seja realizada por especialistas. Assim, todas as análises aqui apresentadas foram balizadas por um conjunto reduzido, classificado e pareado de textos concebido por uma equipe de linguistas liderada na Universidade Federal do Rio Grande do Sul (UFRGS)(FINATTO; TCACENCO, 2021). O *corpus* em questão está representado na Figura 14 pelo conjunto *Museu* e contém 42 textos originais e suas simplificações, totalizando 84 textos. São textos cuja função é acompanhar experimentos expostos em um museu gaúcho de ciências e tecnologia. Os textos originais foram escritos para público leigo em geral, e suas versões simplificadas foram adaptadas lexical e sintaticamente para

alunos do final do Ensino Fundamental de escolas públicas (FINATTO; TCACENCO, 2021).

5.1.2 Enriquecimento de Dados para Teste

Para avaliar se os resultados obtidos a partir das técnicas de AD poderiam ser generalizados para outro domínio, buscaram-se dados de textos originais e simplificados pareados de outros domínios do conhecimento para teste. Dentre os trabalhos identificados no Capítulo 4, o único repositório consolidado deste gênero e ativo foi o disponibilizado por Wilkens, Zilio et al. (2016). Este trabalho proporcionou a consulta ao *corpus Wikibooks*, previamente classificado em relação ao sistema educacional brasileiro. Para conversão em "simples" e "complexo", considerou-se apenas que o primeiro nível (ensino fundamental) conteria os textos simples, e os demais, complexos. Os textos coletados da *web* por Wilkens, Zilio et al. (2016) foram desconsiderados dada a possibilidade de imprecisão da anotação disponibilizada, visto terem sido rotulados por meio de um modelo automático de classificação, o que tenderia à propagação de erros.

Além do *corpus Wikibooks*, foram considerados textos da coleção *Literatura para Todos*, publicada pelo BRASIL (2006) e distribuída para jovens e adultos recém-alfabetizados. Este caso é interessante, pois Rodrigues, Freitas e Quental (2013) ponderam que os seis livros analisados dessa coleção são complexos para o público pretendido pelo estudo realizado pelas autoras. Ressalta-se que o trabalho de Rodrigues, Freitas e Quental (2013) não foi incluído na Tabela 4 porque apesar de utilizarem ferramentas automáticas para balizar seus resultados, a avaliação em si é proposta de forma não automática.

Sobre os textos da camada de enriquecimento, devido à limitação da ferramenta NILC-Metrix de 2000 palavras para processamento e do tempo para processamento de incorporação de palavras para textos muito longos, foram consideradas apenas as primeiras 2000 palavras dos textos para captura da representação de atributos.

5.2 Classificação

Para análise dos conjuntos gerados artificialmente, além de uma ponderação holística em termos qualitativos, considerou-se a tarefa de AAL. Conforme pode ser identificado na Tabela 4, o trabalho de Wilkens, Zilio et al. (2016) é particularmente interessante haja vista que o problema de pesquisa se aproxima do trabalho aqui apresentado. Neste contexto, para fins comparativos, considerou-se o uso de um RL para o treinamento do classifica-

dor automático. Adicionalmente, foram utilizados o SVM e o *XGBoost* identificados nos demais trabalhos apresentados na Tabela 4.

Devido ao baixo número de textos do *corpus* principal, emprega-se o método conhecido por *Leave-one-out (LOO)* para treinamento e validação dos resultados do classificador. O LOO é um caso especial da validação cruzada conhecida por *k-fold*, em que o número de divisões k é igual ao número de instâncias i do conjunto de dados. De acordo com Wong (2015) é indicado o uso do LOO para conjuntos de dados pequenos para se obter uma estimativa de precisão mais confiável em algoritmos de classificação. Desta forma, considerando o *corpus* inicial com $i = 84$ textos, o modelo foi treinado 84 vezes, sendo que cada vez um texto específico era desconsiderado do conjunto de treinamento e usado apenas para avaliar o modelo. Ressalta-se que além do texto em si, o seu par com etiqueta oposta também foi desconsiderado nos treinos, de modo a evitar um vazamento dos textos de teste no treinamento e ocasionar sobre ajustes no classificador. No mesmo sentido, no caso dos treinos com uso de AD, os textos aumentados relacionados ao texto excluído, bem como ao seu par oposto também não eram considerados naquela rodada.

No desenvolvimento do processo de classificação foi utilizado o pacote Scikit-Learn (SCIKIT-LEARN, 2021a,c), e como forma de facilitar a flexibilização de parâmetros dos modelos, optou-se pela criação de um arquivo de configuração do tipo **.ini*, permitindo a execução de inúmeros testes para diferentes combinações de atributos, bem como *corpora* de treino/validação. Sobre a estrutura do arquivo de configuração, consideremos o exemplo a seguir:

```
[loo_rt_simples_ss_museu_bertimbau]
clf_list = logistic_regression,svm,xgb
output_name = loo_rt_simples_ss_museu_bertimbau
output_detail = leave one out: textos do museu aumentados por rt da classe
    simples e ss com embeddings do bertimbau e mtricas do nilc normalizadas
validation_type = loo
corpus_selection_list = SS_Museu,RT_Museu,Museu
augmentation_label = all,Simples,all
type_embeds_cols = bertimbau
type_metrics_cols = none
```

Nesta situação ao se deparar com o bloco `loo_rt_simples_ss_museu_bertimbau`, identifica-se que serão realizadas três iterações de classificação, uma com o modelo de RL, outra com SVM e, por fim, *XGBoost*. O arquivo de saída terá o nome dado pelo

parâmetro `output_name`. Para facilitar a identificação dos modelos, foi acrescentado um campo de detalhamento `output_detail` por conta das inúmeras combinações possíveis de atributos, métricas, validações e seleções de *corpora*. O tipo de validação do modelo, dado por `validation_type` será por LOO e o conjunto de dados selecionado para treino e teste é dado pelos textos aumentados por SS (`SS_Museu`), RT (`RT_Museu`) e os textos originais do museu, conforme o parâmetro `corpus_selection_list`. A ordem da disposição dos itens no `corpus_selection_list`, está relacionada à ordem dos rótulos indicados no parâmetro `augmentation_label`. De modo que, no exemplo, serão usadas todas as etiquetas para os textos originais do museu e aumentados por SS, já no caso dos aumentados por RT será realizado um aumento desbalanceado apenas os textos da classe "simples".

Em termos de entrada para o modelo de AAL, temos como atributos a possibilidade de concatenação das incorporações de palavras e métricas do NILC-Metrix normalizadas. No exemplo acima, as métricas do NILC-Metrix serão ignoradas (conforme indicado pelo parâmetro `type_metrics_cols = none`) e apenas a incorporação de palavras que considera o modelo pré-treinado do *BERTimbau* será utilizada (conforme indicado pelo parâmetro `type_embeds_cols = bertimbau`).

Desta forma, no exemplo acima serão treinados e avaliados três modelos de classificação (por RL, SVM e XGBoost) com incorporações de palavras do modelo pré-treinado BERTimbau. Ressalta-se que não foram realizados ajustes finos em termos de parametrização dos modelos de classificação do pacote *Scikit-Learn*, de modo que se seguiu o modelo padrão da biblioteca. Foi considerada apenas a inicialização dos estados aleatórios dos modelos em 0 (zero), a fim de evitar mudanças nos resultados a cada iteração.

5.3 Representação de Atributos

No que diz respeito à representação de atributos textuais como numéricos (*i.e.*, de modo que sirvam de entrada para os modelos de classificação), trabalhos mais recentes recorreram ao uso de métodos de incorporação de palavras, advindos de modelos como o BERT, como massa de dados para entrada dos classificadores (IMPERIAL, 2021). No corrente trabalho, explora-se tanto o uso das representações por incorporação de palavras quanto dos conjuntos de métricas de análise linguística e psicolinguística disponíveis para o PB dados pelo NILC-Metrix (LEAL; SANCHES DURAN et al., 2021; LEAL; SCARTON et al., 2022) normalizados, considerando também análises combinatórias de ambos os formatos de representação de atributos.

Para a extração das incorporações de palavras são consideradas abordagens estáticas e contextualizadas. No caso das estáticas, foram utilizadas as bibliotecas: (1)*Fasttext* (INC, 2022) com carga do modelo pré-treinado `cc.pt.300.bin` (FACEBOOK, 2018), que faz uso da abordagem *word2vec* por *CBOW*, e (2) *Gensim* (ŘEHŮŘEK, 2022) com carga do modelo pré-treinado `skip_s300` (NILC, 2017), que faz uso de *word2vec* por *skip-gram*. Em ambos os modelos, foram utilizadas incorporações de palavras com 300 dimensões; de modo que, durante a varredura dos textos, para cada texto é atribuído um vetor médio dado pelos vetores de incorporações de palavras gerados. Em termos técnicos, cada vetor médio é armazenado em uma lista de vetores, que posteriormente é convertida em um *DataFrame* a ser combinado com base no elemento *key* com os textos do *DataFrame* da camada bronze.

Em relação às abordagens contextualizadas, foi utilizada a biblioteca *transformers*, disponibilizada pelo grupo *Hugging Face* (WOLF et al., 2020) com carga do modelo pré-treinado BERTimbau denominado `neuralmind/bert-base-portuguese-cased` (SOUZA; NOGUEIRA; LOTUFO, 2020) com 768 dimensões. Destaca-se que neste processo foi utilizada a abordagem conhecida por *feature extraction* para obtenção do vetor médio de representação das incorporações de palavras dados para cada texto.

No *feature extraction*, o modelo pré-treinado é utilizado para gerar as incorporações de palavras de um texto, sendo que as camadas do modelo pré-treinado permanecem inalteradas. Uma outra abordagem que tem sido utilizada e foi previamente mencionada na seção 2.3 é o *fine-tuning*. Neste caso, além de gerar as incorporações de palavras, ajustam-se os pesos do modelo por meio da inclusão de camadas para uma tarefa específica. Esta abordagem geralmente requer mais dados e tempo de treinamento, e optou-se por sua não utilização uma vez que conforme mencionado na seção 3, a análise de Lee, Jang e Lee (2021) demonstra que *conjuntos menores de dados se beneficiam mais do uso de atributos linguísticos do que da incorporação de palavras*. Ou seja, o uso do *fine-tuning* no contexto de leiturabilidade aplicado a um *corpus* reduzido tende ao *overfitting*.

Ainda mais especificamente sobre o modelo *BERTimbau*, há uma restrição de processamento no número de *tokens*, de modo que não é possível converter uma sequência maior do que 512 *tokens* de uma vez. Por isso, foi utilizada uma estratégia de processamento em que o texto é quebrado em sentenças de até X palavras separadas de forma gráfica. Esta abordagem foi considerada por se tratar de um processo menos custoso computacionalmente, evitando o processamento por *tokenização*. No entanto, como elucidado anteriormente, o tamanho de um *token* pode variar em relação a um processo gráfico de

separação de palavras (FREITAS, 2022). Por isso, foi considerada uma margem de cerca de 10% da limitação do número de *tokens* do modelo. Ainda assim, é possível haver inconsistências na geração das incorporações de palavras. Nestes casos, há uma redução de 50 palavras do número X para cada iteração, até que ocorra a existência da representação vetorial para a sentença. Inicialmente, este número X é definido em 450. Ao final do processamento das sentenças, é realizada uma média aritmética entre os vetores gerados para representar cada um dos textos do *corpus*.

No que diz respeito à extração das métricas pelo NILC-Metrix, devido ao número de textos aumentados do *corpus* da camada bronze, torna-se impraticável utilizar a ferramenta *on-line* disponibilizada pelo NILC (LEAL; SCARTON et al., 2022). Desta forma, foi realizada uma clonagem do repositório NILC-Metrix (LEAL, 2022) para execução local. Para conseguir executar o código disponibilizado, foram necessárias algumas adaptações, resumidas a seguir:

- Devido ao tamanho oferecido para os repositórios-padrão no *Github*, o diretório *Tools* não foi disponibilizado. Assim sendo, após contato com o proprietário do repositório, Sidney Leal, foi incluído o endereço para *download* da pasta no arquivo *readme* do repositório. Esta pasta contém todas as bases para rodar ferramentas de terceiros, como *parsers* e *taggers*.
- O acesso ao *parser* PALAVRAS pelo NILC-Metrix considera um servidor pago disponibilizado na rede da Universidade de São Paulo (USP). Por isto, foi realizada uma adaptação local no método palavras flat, conforme indicado no trecho de código abaixo:

```
def palavras_flat(t):
    '''
    Call a webservice to run the parser Palavras

    :param text: the text to be parsed, in unicode.
    :return: the response string from Palavras
    '''
    params = {'text': t.raw_content, 'parser': 'dep-eb', 'visual':
        'plain', 'heads': 'symbol', 'multisearch':'searchtype',
        'inputlang': 'pt'}
    f = requests.get('https://visl.sdu.dk/cgi-bin/visl.pt.cgi',
        params)
```

```
        return f.text.replace('\\n', '\n').replace('\\t',
            '\t').replace('', 's')
```

A adaptação considerou a mudança da requisição para acesso direto à *Application Programming Interface* (API) do PALAVRAS, após investigação do funcionamento da página em que o recurso está hospedado (UNIVERSITY, 2022).

- Identificou-se que o NILC-Metrix é executado com base em três imagens [1]:

  1. *postgres:latest*: esta é a imagem que contém o banco de dados para rodar a aplicação;
  2. *cohmetrix:focal*: esta imagem gera o *container* que é responsável por executar a aplicação em si;
  3. *nilcmetrix:nilcmetrix*: esta imagem está relacionada à parte final da aplicação, *i.e.*, da camada de conexão *web*.

Por isso, foi necessário realizar a instalação do programa Docker[2]. Para acesso à camada de execução, uma vez inicializados os *containers*, foi vinculado o acesso do *container* gerado pela imagem *cohmetrix:focal* ao programa *Visual Studio*. Isto permitiu a execução de um script *Python* iterando os textos do *corpus* da camada bronze para acesso direto à biblioteca *text_ metrics*, base para a ferramenta NILC-Metrix, conforme trecho de código abaixo.

```
import pandas as pd
import text_metrics
import datetime
import os

''' Identifica textos que ja tenham sido processados'''
prev_keys = []
for filepath, _, files in os.walk(os.getcwd() + "/parquets/"):
    for item in files:
        prev_keys.append(str(item)[:-8])
```

[1] Neste contexto, uma imagem é um modelo para um *container*. Ou seja, a partir de imagens é possível criar *containers*. Um *container*, por sua vez, contém todos os componentes necessários para executar uma aplicação, incluindo o código, as bibliotecas e os arquivos de configuração.

[2] O Docker é um sistema de gerenciamento de *containers* que permite desenvolver e executar aplicações em *containers*.

```
''' Carrega textos para processamento '''
bronze = pd.read_parquet('bronze.parquet')

''' Itera textos para processamento '''
for i, item in bronze.iterrows():
    if not(item['key'] in prev_keys):
        print(f"Processando {item['key']} -
            {datetime.datetime.now()}")

        ''' Captura metricas para o texto '''
        t = text_metrics.Text(item['Content'])
        ret =
            text_metrics.nilc_metrics.values_for_text(t).as_flat_dict()

        ''' Converte metricas para o formato de DataFrame '''
        df = pd.DataFrame(ret, index=[i])
        df['key'] = item['key']

        ''' Armazena informacao '''
        df.to_parquet(os.getcwd() + "/parquets/" + item['key'] +
            ".parquet")

        print(f"Encerrado {i} de {len(bronze['key'])} -
            {datetime.datetime.now()}")
```

- Observa-se que no código acima, foi criada uma lista denominada *prev_keys* apenas para evitar reprocessamento de textos já processados caso seja necessária reexecução do código. Isto foi necessário porque foram identificadas intermitências na execução da aplicação relacionadas ao acesso de ferramentas de terceiros, o que acaba por gerar interrupções e necessidade de reexecução do código.

Ainda sobre as métricas do NILC-Metrix, é intuitivo avaliar o uso de seleção de atributos perante os agrupamentos apresentados na seção 2.3.2. Destaca-se, no entanto, que foram realizadas implementações preliminares com esta abordagem para textos completos, e nas seleções de atributos criadas (supervisionadas ou não), os resultados foram inferiores ao uso das 200 métricas existentes na ferramenta. Este comportamento foi analogamente observado na literatura em trabalhos anteriores (SCHWARM; OSTENDORF,

2005; SCARTON; GASPERIN; ALUISIO, 2010; ALUISIO et al., 2010; WAGNER FILHO; WILKENS; VILLAVICENCIO, 2016; GAZZOLA; LEAL; ALUISIO, 2019). Assim, optou-se na metodologia aqui proposta, por não seguir com a inclusão de métodos de seleção de atributos para as métricas do NILC-Metrix.

6 Resultados Experimentais

Para avaliar empiricamente as técnicas de AD propostas neste trabalho, foram consideradas 75 combinações de métodos de classificadores, representações de atributos, e conjuntos de textos para treinamento de diferentes modelos supervisionados de classificação por leiturabilidade, voltados para AAL. Devido ao número de possibilidades para análise, necessidade de modelagem dos dados de saída, bem como de variações no nível de detalhamento (do inglês, *drill down* e *drill up*) dos dados para proporcionar análise tanto macro quanto mais específicas, optou-se por utilizar a ferramenta *Power BI* para visualização.

Para cada uma das saídas dos classificadores treinados e testados com diferentes entradas, uma linha é indicada nos painéis contidos nas Figuras 15, 16, 17, 18, 19 e 20. Dado que:

- A coluna `CLF` indica o tipo da técnica utilizada para o AM da classificação;
- A coluna `TV` indica o tipo de validação utilizado - existem duas opções: `selection` ou `loo`. A primeira é um processo simples em que os *corpora* de treino e teste diferem totalmente entre si. Por isso, foi utilizada nos casos de teste por enriquecimento de dados, em que o treino se mantém com o *corpus* principal, mas o teste é realizado com *corpora* de domínios distintos. Já a segunda é referente ao processo de LOO, detalhado na seção 5.2;
- A coluna `Sel.Embeds` indica o modelo que foi utilizado para coletar as incorporações de palavras. Quando `Sel.Embeds` = *none* significa que nenhum modelo de incorporação de palavras foi atribuído ao modelo, o oposto ocorre quando `Sel.Embeds` = *all*. Também é possível selecionar apenas atributos dos modelos indicados na seção 5.3 de forma individualizada, de modo que, `Sel.Embeds` = *bertimbau* é usado para indicar que o modelo considerou como entrada as incorporações de palavras contextualizadas geradas pelo modelo pré-treinado BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020); `Sel.Embeds` = fb_cbow é usado para indicar que o modelo con-

siderou como entrada as incorporações de palavras estáticas geradas pelo modelo pré-treinado *Fasttext* (FACEBOOK, 2018); e `Sel.Embeds` = *nilc_sg* é usado para indicar que o modelo considerou como entrada as incorporações de palavras estáticas geradas pelo modelo pré-treinado `skip_s300` (NILC, 2017).

- A coluna `Sel.Métricas` indica a seleção das métricas de coesão e coerência textuais advindas do *NILC-Metrix*. Assim, nas figuras a seguir, quando `Sel.Métricas` = *none* significa que nenhuma das métricas de coesão e coerência textuais advindas do *NILC-Metrix* (LEAL; SCARTON et al., 2022) são utilizadas como entrada para o modelo; enquanto, quando `Sel.Métricas` = *all*, o oposto é verdadeiro. Ressalta-se que, embora, neste trabalho esta seleção seja binária, o formato de implementação criado, permite que sejam selecionados grupos de métricas específicos dentre as 200 métricas disponibilizadas pelo *NILC-Metrix*[1].

- O nome da coluna `Corpora de Treino` é indicativo por si só, no entanto, ressalta-se que foram realizadas diferentes combinações de modo que os nomes dos *corpora* incluídos para treino estão separados por vírgulas e são representados por:

  - Museu: *corpus* principal;
  - SS_Museu: *corpus* com textos aumentados por SS de forma balanceada, *i.e.*, para ambas as classes (simples e complexo);
  - SS_Museu_S: *corpus* com textos aumentados por SS de forma desbalanceada, apenas para a classe simples;
  - SS_Museu_C: *corpus* com textos aumentados por SS de forma desbalanceada, apenas para a classe complexa;
  - RT_Museu: textos aumentados por RT de forma balanceada, *i.e.*, para ambas as classes (simples e complexo);
  - RT_Museu_S: *corpus* com textos aumentados por RT de forma desbalanceada, apenas para a classe simples;
  - RT_Museu_C: *corpus* com textos aumentados por RT de forma desbalanceada, apenas para a classe complexa;

- Com uso da técnica de matriz de confusão, as colunas TP, TN, FN e FP compõem:

[1] Conforme mencionado na seção 5.3, os resultados preliminares de tais seleções foram inferiores ao uso completo das métricas da ferramenta. Assim, optou-se por não seguir com a inclusão de métodos de seleção de atributos para as métricas do NILC-Metrix neste trabalho.

  - TP (*True Positive*): o número de textos que foram classificados pelo modelo como verdadeiros positivos, *i.e.*, o rótulo do texto em validação era simples e o modelo o classificou corretamente como simples.
  - TN (*True Negative*): o número de textos que foram classificados pelo modelo como verdadeiros negativos, *i.e.*, o rótulo do texto em validação era complexo e o modelo o classificou corretamente como complexo.
  - FN (*False Negative*): o número de textos que foram classificados pelo modelo como falsos negativos, *i.e.*, o rótulo do texto em validação era simples e o modelo o classificou como complexo.
  - FP (*False Positive*): o número de textos que foram classificados pelo modelo como falsos positivos, *i.e.*, o rótulo do texto em validação era complexo e o modelo o classificou como simples.
  - Total: número total de textos usados na validação do modelo.

- Para uma visão ponderada em relação às quantidades acima, foram consideradas as seguintes métricas:
  - *Acurácia* dada por $([TP] + [TN])/[Total]$
  - *F1-Score*, média entre $2 * [Precisao] * [Revocacao]/([Precisao] + [Revocacao])$ de ambas as classes, em que Precisão $= [TP]/([TP] + [FP])$ e Revocação $= [TP]/([TP] + [FN])$

Considerou-se a ordenação decrescente dos painéis apresentados na sequência pela coluna de acurácia, uma vez que, no geral, os conjuntos de dados utilizados estão balanceados e não há uma preferência na classificação dos textos simples ou complexos.

6.1 Avaliação do *corpus* principal

Numa visão inicial com testes para os modelos sem uso de AD, *i.e.*, treinados e testados apenas com os textos originais do *corpus* principal, os modelos que apresentaram melhor resultado fizeram uso da validação por LOO e do classificador do tipo RL, ambos alcançando uma acurácia de 94,0%, vide Figura 15. A diferença entre os dois é que o primeiro considera apenas a representação de atributos por incorporação de palavras treinada com o modelo do *BERTimbau* e o segundo considera tanto os atributos advindos do *BERTimbau* quanto das métricas do *NILC-Metrix*. A escolha desta segunda combinação se deu

por conta das avaliações individuais das representações de atributos, em que foi observado que o modelo de RL apenas com as incorporações de palavras do *BERTimbau* e o modelo por RL apenas com as métricas do *NILC-Metrix* se sobressaíram em relação aos modelos treinados apenas com as incorporações de palavras estáticas. Ao realizar a combinação, notou-se que o modelo atingiu o mesmo resultado daquele usando apenas as incorporações de palavras do *BERTimbau*.

Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
Museu	Museu	rl	loo	bertimbau	none	40	39	2	3	84	94,0%	94,0%
Museu	Museu	rl	loo	bertimbau	all	40	39	2	3	84	94,0%	94,0%
Museu	Museu	svm	loo	all	all	37	41	5	1	84	92,9%	92,8%
Museu	Museu	svm	loo	bertimbau	none	38	39	4	3	84	91,7%	91,7%
Museu	Museu	rl	loo	all	all	38	39	4	3	84	91,7%	91,7%
Museu	Museu	rl	loo	fb_cbow	none	34	38	8	4	84	85,7%	85,7%
Museu	Museu	xgb	loo	all	all	35	37	7	5	84	85,7%	85,7%
Museu	Museu	rl	loo	none	all	37	34	5	8	84	84,5%	84,5%
Museu	Museu	svm	loo	none	all	35	36	7	6	84	84,5%	84,5%
Museu	Museu	xgb	loo	none	all	35	36	7	6	84	84,5%	84,5%
Museu	Museu	xgb	loo	bertimbau	none	35	32	7	10	84	79,8%	79,7%
Museu	Museu	rl	loo	nilc_sg	none	32	35	10	7	84	79,8%	79,7%
Museu	Museu	svm	loo	fb_cbow	none	30	34	12	8	84	76,2%	76,1%
Museu	Museu	xgb	loo	fb_cbow	none	34	30	8	12	84	76,2%	76,1%
Museu	Museu	svm	loo	nilc_sg	none	25	37	17	5	84	73,8%	73,3%
Museu	Museu	xgb	loo	nilc_sg	none	28	32	14	10	84	71,4%	71,4%

Figura 15: Painel com os resultados para classificadores considerando apenas o *corpus* principal para treino e teste, com diferentes combinações do tipo do classificador e escolhas das representações de atributos

Analisando esses primeiros resultados, podemos pontuar que:

PROPOSIÇÃO 1

O conceito de leiturabilidade se correlaciona ao processo de predição probabilística dada uma vizinhança de palavras, conforme supunha Taylor (1953), haja vista que os classificadores com melhores resultados foram aqueles treinados com uso de incorporações de palavras contextualizadas.

Devido ao fato do RL ter apresentado resultados significativamente superiores que os demais classificadores, e ainda, de que o potencial de crescimento do número de combinações tende a aumentar significativamente a medida em que se aprofunda na análise exploratória, as próximas análises foram executados apenas com classificadores por RL. Da mesma forma, a escolha da representação de atributos de entrada foi restringida para: (1) uso das incorporações de palavras por *BERTimbau* e (2) combinação das incorporações

de palavras por *BERTimbau* e métricas do *NILC-Metrix*.

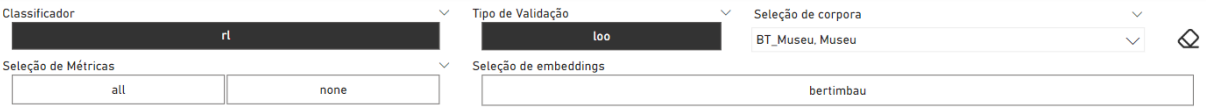
Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
Museu	SS_Museu, Museu	rl	loo	bertimbau	all	40	39	2	3	84	94,0%	94,0%
Museu	SS_Museu_C, Museu	rl	loo	bertimbau	all	40	39	2	3	84	94,0%	94,0%
Museu	SS_Museu_S, Museu	rl	loo	bertimbau	all	40	39	2	3	84	94,0%	94,0%
Museu	SS_Museu, Museu	rl	loo	bertimbau	none	39	38	3	4	84	91,7%	91,7%
Museu	SS_Museu_C, Museu	rl	loo	bertimbau	none	39	38	3	4	84	91,7%	91,7%
Museu	SS_Museu_S, Museu	rl	loo	bertimbau	none	38	39	4	3	84	91,7%	91,7%
Museu	SS_Museu, Museu	rl	loo	none	all	37	34	5	8	84	84,5%	84,5%

Figura 16: Painel com os resultados para classificadores considerando treino com *corpus* principal e *corpus* aumentado por SS

No caso dos treinos com os dados aumentados por SS, foram considerados aumentos tanto desbalanceados quanto balanceados conforme representado na Figura 16. Observa-se que o uso dos dados artificiais não gerou ganhos ou perdas significativas em relação aos modelos treinados apenas com os textos sem aumento. A partir desta verificação, podemos concluir que:

PROPOSIÇÃO 2

Apesar da SS ser uma ferramenta de apoio à tarefa de ST, acrescenta pouca informação de padrões linguísticos em termos de aprendizado no âmbito da representação por incorporação de palavras. Isto porque as incorporações de palavras com contextos semelhantes se aproximam. Ou seja, há pouca variação entre as representações do texto original e aumentados, haja vista que sinônimos são posicionados muito próximos uns aos outros no espaço latente.



Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
Museu	RT_Museu_S, Museu	rl	loo	bertimbau	all	39	39	3	3	84	92,9%	92,9%
Museu	RT_Museu, Museu	rl	loo	bertimbau	none	39	38	3	4	84	91,7%	91,7%
Museu	RT_Museu, Museu	rl	loo	bertimbau	all	39	37	3	5	84	90,5%	90,5%
Museu	RT_Museu_C, Museu	rl	loo	bertimbau	all	37	39	5	3	84	90,5%	90,5%
Museu	RT_Museu_S, Museu	rl	loo	bertimbau	none	40	35	2	7	84	89,3%	89,2%
Museu	RT_Museu_C, Museu	rl	loo	bertimbau	none	34	39	8	3	84	86,9%	86,9%

Figura 17: Painel com os resultados para classificadores considerando treino com *corpus* principal e *corpus* e aumentado por RT

Em relação aos treinos realizados com os dados aumentados por RT, os resultados são

apresentados na Figura 17. A expectativa relacionada ao uso de aumento por RT era de que o método pudesse impulsionar os resultados por causar alterações mais significativas na estrutura do texto. Entretanto, a Figura 17 demonstra uma perda de acurácia a despeito das combinações de atributos. Observa-se, todavia, que o modelo treinado com os textos originais do *corpus* principal e os aumentados por RT apenas para a classe simples se sobressaíram em relação aos demais modelos da Figura 17. Esta consideração nos leva a outro resultado importante:

PROPOSIÇÃO 3

O processo de RT acaba atuando como uma ferramenta de apoio à ST. Isto significa que o aumento da classe complexa a partir do uso de RT gerou textos de rótulos duvidosos, o que acabou por prejudicar o aprendizado do classificador.

Ao avaliar os resultados das Figuras 16 e 17, é possível observar que os classificadores treinados com textos aumentados[2], ainda que de forma pouco significativa, obtiveram melhores resultados com a inclusão das métricas do *NILC-Metrix*. Sendo assim, infere-se que:

PROPOSIÇÃO 4

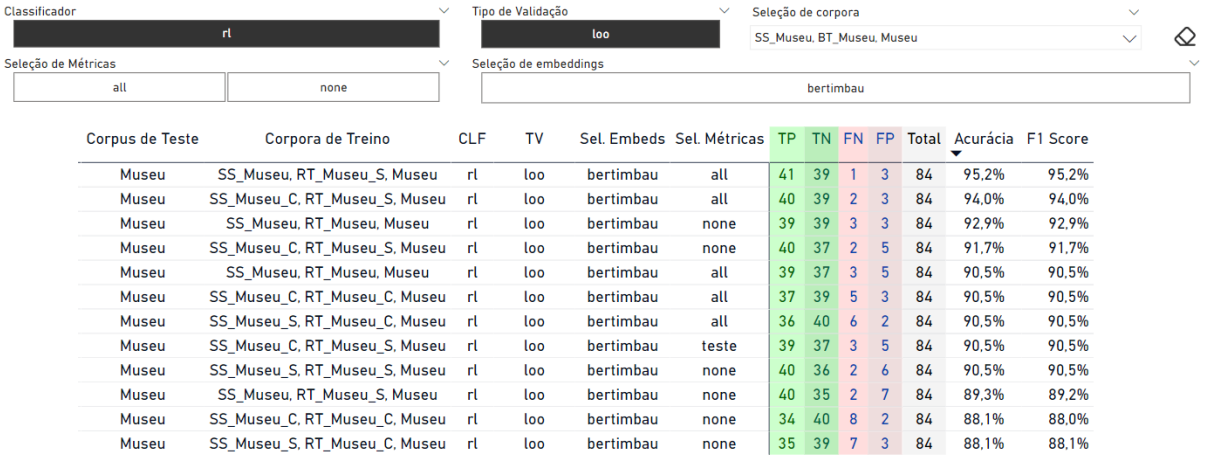
Embora as incorporações de palavras contextualizadas se apresentem como um forte indicativo para a leiturabilidade, o uso de métricas dados pela ferramenta *NILC-Metrix* agrega informações que não são consideradas nas incorporações de palavras.

Na sequência, para avaliação em termos de combinações dos aumentos, optou-se por:

- Treino com textos do *corpus* principal e aumentados para ambas as classes tanto por SS quanto por RT.
- Treino com textos do *corpus* principal e aumentados apenas para a classe simples por RT e da classe complexa por SS. Esta opção foi uma tentativa de balancear o aumento, considerando que o RT apresentou resultados significativamente melhores com aumento apenas da classe simples.
- Treino com textos do *corpus* principal e aumentados apenas para a classe simples por RT e de ambas as classes por SS. Esta opção foi uma tentativa de considerar melhores

[2]Com exceção do aumento por RT da classe complexa, que apresenta questões relacionadas à incerteza da manutenção dos rótulos dos textos originais, conforme indicado na proposição 3.

resultados de ambos os aumentos sem causar um desbalanceamento expressivo no treino.



Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
Museu	SS_Museu, RT_Museu_S, Museu	rl	loo	bertimbau	all	41	39	1	3	84	95,2%	95,2%
Museu	SS_Museu_C, RT_Museu_S, Museu	rl	loo	bertimbau	all	40	39	2	3	84	94,0%	94,0%
Museu	SS_Museu, RT_Museu, Museu	rl	loo	bertimbau	none	39	39	3	3	84	92,9%	92,9%
Museu	SS_Museu_C, RT_Museu_S, Museu	rl	loo	bertimbau	none	40	37	2	5	84	91,7%	91,7%
Museu	SS_Museu, RT_Museu, Museu	rl	loo	bertimbau	all	39	37	3	5	84	90,5%	90,5%
Museu	SS_Museu_C, RT_Museu_C, Museu	rl	loo	bertimbau	all	37	39	5	3	84	90,5%	90,5%
Museu	SS_Museu_S, RT_Museu_C, Museu	rl	loo	bertimbau	all	36	40	6	2	84	90,5%	90,5%
Museu	SS_Museu_C, RT_Museu_S, Museu	rl	loo	bertimbau	teste	39	37	3	5	84	90,5%	90,5%
Museu	SS_Museu_S, RT_Museu_S, Museu	rl	loo	bertimbau	none	40	36	2	6	84	90,5%	90,5%
Museu	SS_Museu, RT_Museu_S, Museu	rl	loo	bertimbau	none	40	35	2	7	84	89,3%	89,2%
Museu	SS_Museu_C, RT_Museu_C, Museu	rl	loo	bertimbau	none	34	40	8	2	84	88,1%	88,0%
Museu	SS_Museu_S, RT_Museu_C, Museu	rl	loo	bertimbau	none	35	39	7	3	84	88,1%	88,1%

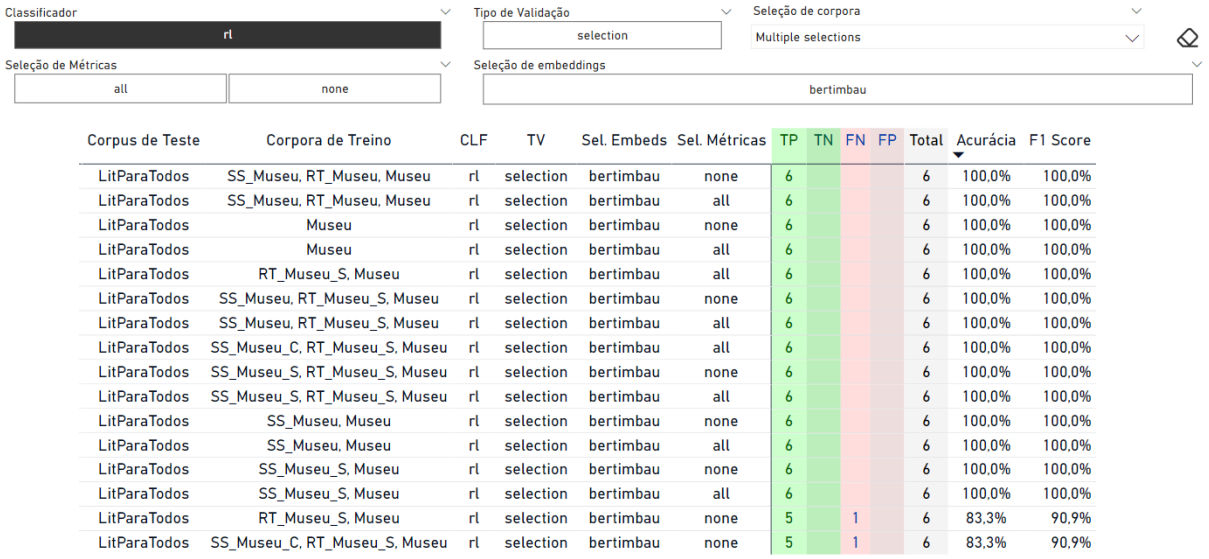
Figura 18: Painel com os resultados para classificadores considerando treino com *corpus* principal e *corpus* aumentado por combinações de RT e SS

Observa-se que por meio da combinação de SS balanceada com RT apenas para a classe simples foi possível melhorar o resultado do classificador. No entanto, na próxima seção foram utilizadas diversas combinações para avaliação da generalização dos modelos quando aplicáveis a outros domínios, haja vista que os resultados das combinações não apresentaram uma variância significativa a ponto de justificar uma preferência combinacional.

6.2 Incorporação de *corpus* de outros domínios

No que diz respeito ao uso de dados de outros domínios, foram considerados os textos do *Wikibooks* (WILKENS; ZILIO et al., 2016) e da coleção *Literatura para Todos* (RODRIGUES; FREITAS; QUENTAL, 2013).

Em relação à coleção *Literatura para Todos* (RODRIGUES; FREITAS; QUENTAL, 2013), a análise de Rodrigues, Freitas e Quental (2013) aponta para uma complexidade elevada no que tange à leiturabilidade para neoleitores, *i.e.*, grupo de jovens e adultos recém-alfabetizados; observa-se pela Figura 19 que os seis textos avaliados pelos classificadores treinados com o *corpus* principal, independentemente de aumentado ou não, foram classificados como simples. Com exceção do texto *Tubarão* para o caso do modelo treinado com *corpus* principal e *corpus* aumentado por RT para a classe simples, tanto com e sem adição de SS para a classe complexa considerando como atributos apenas as



Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
LitParaTodos	SS_Museu, RT_Museu, Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	SS_Museu, RT_Museu, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	RT_Museu_S, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	SS_Museu, RT_Museu_S, Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	SS_Museu, RT_Museu_S, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	SS_Museu_C, RT_Museu_S, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	SS_Museu_S, RT_Museu_S, Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	SS_Museu_S, RT_Museu_S, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	SS_Museu, Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	SS_Museu, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	SS_Museu_S, Museu	rl	selection	bertimbau	none	6				6	100,0%	100,0%
LitParaTodos	SS_Museu_S, Museu	rl	selection	bertimbau	all	6				6	100,0%	100,0%
LitParaTodos	RT_Museu_S, Museu	rl	selection	bertimbau	none	5		1		6	83,3%	90,9%
LitParaTodos	SS_Museu_C, RT_Museu_S, Museu	rl	selection	bertimbau	none	5		1		6	83,3%	90,9%

Figura 19: Painel com os resultados para classificadores considerando treino com aumentos ou não do *corpus* principal e teste com textos da coleção *Literatura para Todos*

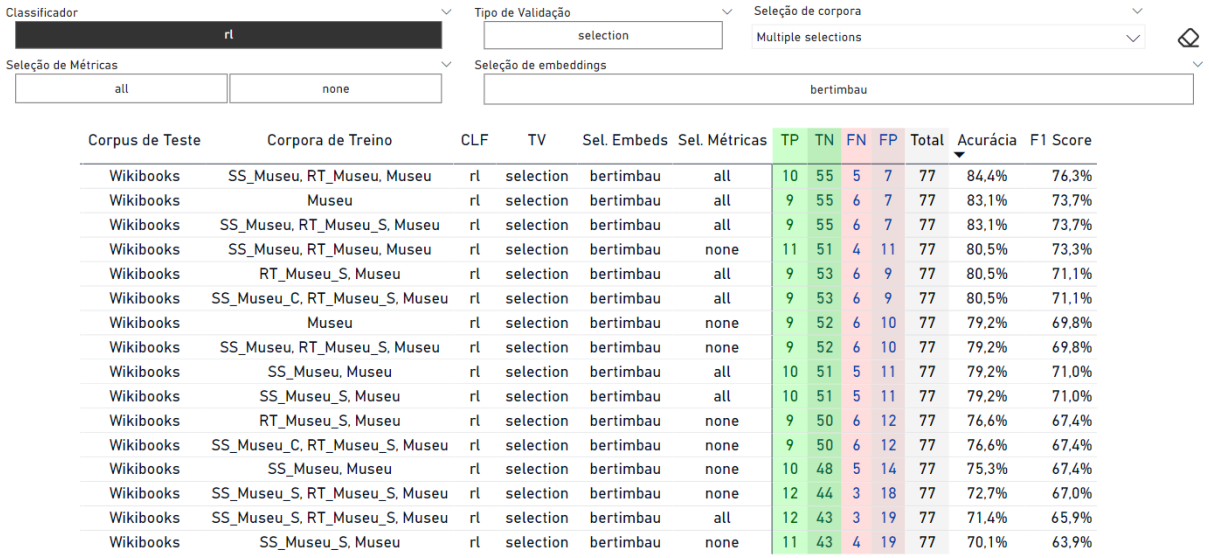
incorporações de palavras do *BERTimbau*. Ressalta-se que este texto é considerado um dos mais complexos pela análise de Rodrigues, Freitas e Quental (2013), juntamente do texto *Léo, o pardo*. Aqui é importante destacar também que os classificadores da Figura 19 foram treinados sob a ótica dos textos do *corpus* principal, cujas simplificações foram destinadas ao público escolar infanto-juvenil e não ao público de neoleitores (mais especificamente, de adultos do EJA). Assim, é esperado que haja um relaxamento maior em termos de classificação dos modelos aqui propostos. Neste sentido, podemos afirmar que:

PROPOSIÇÃO 5

A leiturabilidade é um conceito altamente subjetivo com muitos graus de avaliações. A fim de minimizar os impactos desta especificidade, o ideal é que sejam realizados testes envolvendo especialistas e o público leitor-alvo a fim de que a escolha dos textos balizadores de um modelo de classificação, independentemente de automatizado ou não, atendam ao propósito da acessibilidade textual.

Quanto à análise dos resultados dos testes para a coleção *Wikibooks*, a combinação de textos aumentados, incorporações de palavras do BERTimbau e métricas do NILC-Metrix, aumentou a acurácia do classificador para 84,4%. Neste contexto, é importante mencionar que o RL de Wilkens, Zilio et al. (2016) foi elaborado para três classes de modo que a média *F-Score* do RL foi de 0.691. No entanto, uma vez que Wilkens, Zilio et al. (2016) disponibilizou o *F-Score* para cada uma das classes individualmente, é possível

comparar a classificação proposta na metodologia aqui apresentada com o *F-Score* da classe do nível 1, que foi de 0.741 (e o maior entre os três níveis). Destaca-se também que, para o teste aqui proposto, o nível 1 foi definido como classe "simples", e os demais níveis do *corpus*, *i.e.*, 2 e 3, como parte da classe "complexa". Assim, faz sentido a comparação ser realizada com o melhor cenário proposto por Wilkens, Zilio et al. (2016), embora os classificadores tenham sido treinados com um número de classes distintos. Os resultados obtidos para tais testes podem ser observados na Figura 20.



Corpus de Teste	Corpora de Treino	CLF	TV	Sel. Embeds	Sel. Métricas	TP	TN	FN	FP	Total	Acurácia	F1 Score
Wikibooks	SS_Museu, RT_Museu, Museu	rl	selection	bertimbau	all	10	55	5	7	77	84,4%	76,3%
Wikibooks	Museu	rl	selection	bertimbau	all	9	55	6	7	77	83,1%	73,7%
Wikibooks	SS_Museu, RT_Museu_S, Museu	rl	selection	bertimbau	all	9	55	6	7	77	83,1%	73,7%
Wikibooks	SS_Museu, RT_Museu, Museu	rl	selection	bertimbau	none	11	51	4	11	77	80,5%	73,3%
Wikibooks	RT_Museu_S, Museu	rl	selection	bertimbau	all	9	53	6	9	77	80,5%	71,1%
Wikibooks	SS_Museu_C, RT_Museu_S, Museu	rl	selection	bertimbau	all	9	53	6	9	77	80,5%	71,1%
Wikibooks	Museu	rl	selection	bertimbau	none	9	52	6	10	77	79,2%	69,8%
Wikibooks	SS_Museu, RT_Museu_S, Museu	rl	selection	bertimbau	none	9	52	6	10	77	79,2%	69,8%
Wikibooks	SS_Museu, Museu	rl	selection	bertimbau	all	10	51	5	11	77	79,2%	71,0%
Wikibooks	SS_Museu_S, Museu	rl	selection	bertimbau	all	10	51	5	11	77	79,2%	71,0%
Wikibooks	RT_Museu_S, Museu	rl	selection	bertimbau	none	9	50	6	12	77	76,6%	67,4%
Wikibooks	SS_Museu_C, RT_Museu_S, Museu	rl	selection	bertimbau	none	9	50	6	12	77	76,6%	67,4%
Wikibooks	SS_Museu, Museu	rl	selection	bertimbau	none	10	48	5	14	77	75,3%	67,4%
Wikibooks	SS_Museu_S, RT_Museu_S, Museu	rl	selection	bertimbau	none	12	44	3	18	77	72,7%	67,0%
Wikibooks	SS_Museu_S, RT_Museu_S, Museu	rl	selection	bertimbau	all	12	43	3	19	77	71,4%	65,9%
Wikibooks	SS_Museu_S, Museu	rl	selection	bertimbau	none	11	43	4	19	77	70,1%	63,9%

Figura 20: Painel com os resultados para classificadores considerando teste com textos do *Wikibooks*

Uma vez encerradas as combinações exploratórias propostas, é possível responder às questões de pesquisa elencadas na seção 1.2:

- *Qual o efeito da inclusão de exemplos gerados artificialmente para o processo de classificação por AAL?* A inclusão de textos aumentados por meio da combinação dos métodos de SS e RT gerou resultados superiores do que quando usadas individualmente no treinamento de modelos classificatórios para AAL.

- *Quais atributos de entrada de um modelo de AAL por AM produzem resultados mais aderentes com o problema da classificação por leiturabilidade?* Na média, os modelos treinados com atributos combinados, *i.e.*, incorporações de palavras do BERTimbau e métricas do NILC-Metrix, apresentaram resultados superiores àqueles treinados com atributos segregados.

- *Considerando um domínio específico D_w com poucos exemplos de textos simplificados, e um classificador C_k para decidir se um texto é 'simples' ou não, construído a*

partir de dados de domínios genéricos $D_g = \{D_1, \cdots, D_n\}$ *com uso de métodos de AD.* C_k *pode ser generalizável para o domínio específico* $D_w \notin D_g$*?* Sim, o uso dos métodos de AD propostos gerou uma maior generalização do modelo ao considerar diferentes domínios, garantindo melhores resultados para os *corpora* advindos da camada de enriquecimento de dados.

Desta forma, entende-se que a investigação realizada foi capaz de corroborar a hipótese de que os métodos de AD podem reduzir o gargalo de informação em modelos de AAL para PB. E, ao mesmo tempo, disponibilizou via GitHub[3] *corpora* classificados para a tarefa de AAL e a implementação dos métodos de SS e RT para o PB, bem como de um classificador binário textual automático para leiturabilidade.

[3] https://github.com/MeLLL-UFF/text-simplification/tree/dissertacao_luiza-menezes/aumento_de_dados_classificacao

7 Considerações Finais

Nesta dissertação, foram desenvolvidos dois métodos de AD: SS e RT, avaliados por meio de diferentes combinações de métodos, seleção de atributos e de conjuntos de textos no treinamento e teste de modelos supervisionados de classificação por leiturabilidade, com o objetivo de identificar empiricamente a validade de tais métodos no contexto de AAL.

Em termos holísticos, verificou-se que os textos aumentados por RT apresentaram maior nível de manutenção da estrutura gramatical do que aqueles aumentados por SS. Por outro lado, no que diz respeito às avaliações quantitativas, destaca-se que o aumento individual por SS promoveu melhores resultados que o RT, mas não superiores à linha de base dada pelo modelo não aumentado. Em termos da linha de base, o melhor resultado obtido para o conjunto inicial dos textos do *corpus* principal foi uma taxa de acerto de 94,0% considerando um modelo de RL com entrada apenas por incorporação de palavras contextualizadas. Este resultado foi melhorado para 95,2% no modelo treinado com aumento combinado por RT para a classe simples e SS para ambas as classes, com uma concatenação das métricas do *NILC-Metrix* e da incorporação de palavras contextualizadas como atributos de entrada.

7.1 Limitações e ameaças à validade da metodologia proposta

Nesta análise, observou-se que o limiar que define os agrupamentos dos modelos desenvolvidos é altamente variável, de modo que pequenas alterações na seleção de atributos ou dos textos utilizados para treinamento, impactam significativamente o resultado final. Ressalta-se, que devido ao baixo número de textos considerados neste experimento, o viés de tal análise tem relação direta com a volumetria dos dados.

Destaca-se também que o resultado desta análise oferece ao leitor uma base para que pessoas especialistas consigam desenvolver novos estudos a partir de questionamentos em temas mais específicos, como por exemplo: *"qual o impacto da utilização de outras línguas intermediárias no processo de RT na classificação por leiturabilidade? Será que o resultado*

vai tender à ST?"; *"e se for utilizado um número maior ou menor de substituições para a roleta de SS? Será que estes textos sintéticos beneficiarão a classificação?"*; *"quais os resultados para a tarefa de AAL se for utilizado fine-tuning?"*. Neste sentido, ainda existe uma variedade de temas a serem analisados e que merecem especial atenção.

7.2 Trabalhos Futuros

Esta dissertação tem como foco a experimentação de uma área até então inexplorada na literatura, que é a de AD para classificação de leiturabilidade, mais especificamente para o PB. Considerando os pontos abordados na seção 7.1, é importante que novos trabalhos busquem a otimização e aperfeiçoamentos das técnicas expostas de modo a aprofundar as análises apresentadas de forma mais específica.

É interessante também que sejam explorados como *corpus* de treino, textos com um maior espectro de variações dos padrões lexicais e sintáticos para maior generalização, o que pode ser introduzido via métodos de AD. Por isso, trabalhos futuros podem considerar a inclusão de outros textos no conjunto de treinamento não apenas para ampliar o espectro léxico-sintático, como para permitir a utilização de técnicas como *fine-tuning*, reduzindo as chances de *overfitting*.

Adicionalmente, uma das grandes motivações deste trabalho mencionada no Capítulo 1, é a acessibilidade textual. Entende-se que por meio dos desenvolvimento realizados, é possível ter acesso a um *corpus*, embora reduzido, com alta qualidade de anotação em termos de leiturabilidade, de forma a facilitar novas explorações para a automatização de tarefas por AM, como a ST.

7.3 Conclusões

Em comparações a trabalhos anteriores, ao avaliarmos textos em idiomas de menos recursos, os resultados obtidos foram significativamente superiores aos indicados por Imperial (2021) (para a língua filipina). Tal fato, valida a argumentação de Imperial (2021) de que o conhecimento implicitamente codificado nas incorporações de palavras contextualizadas pode ser usado como um conjunto para idiomas com recursos de baixa volumetria no que tange a leiturabilidade. Além disso, este estudo permite corroborar, para o PB, o conceito de que a leiturabilidade se correlaciona ao processo de predição probabilística dada por uma vizinhança de palavras, conforme supunha Taylor (1953), visto que os classificadores

com melhores resultados foram aqueles treinados com uso de incorporações de palavras contextualizados.

Quando avaliamos outros trabalhos no PB, no contexto de avaliação de diferentes domínios para teste, foi possível validar o modelo treinado com os textos do *corpus* principal com o mesmo *corpus* utilizado e desenvolvido no trabalho de Wilkens, Zilio et al. (2016), o *Wikibooks*. O uso da combinação de textos aumentados, incorporações de palavras do BERTimbau e métricas do NILC-Metrix, aqui proposto, elevou a acurácia do classificador para 84,4%, gerando um aumento quando comparado ao classificador sem aumento. Assim sendo, podemos afirmar que o treino do classificador por meio de um conjunto reduzido avaliado por especialistas em leiturabilidade e aumentados por técnicas simples de AD, promoveu resultados superiores e mais controlados quando comparados ao uso de técnicas massivas como a WaC proposta em Wilkens, Zilio et al. (2016).

Conclui-se, portanto, que, a assertividade de um modelo de segregação de textos por avaliação de leiturabilidade está diretamente relacionada ao conjunto de textos escolhido para treino e que o uso de métodos de AD aplicados a *corpora* de alta qualidade tende a promover resultados positivos. Neste sentido, espera-se que este trabalho possa instigar os leitores a buscar por diferentes combinações no uso de métodos de AD e, ao mesmo tempo, apoiá-los em tarefas de uso avançado de PLN, como a ST, ao disponibilizar um *corpus* pareado e especializado em termos de leiturabilidade para o PB.

7.4 Publicações

MENEZES, L.C.; PAES, A.; FINATTO, M.J.B. Abordagem baseada em Aumento de Dados para Avaliação Automática de Leiturabilidade. Domínios de Lingu@gem (ISSN: 1980-5799 – Qualis A1 no novo Qualis), 2023

MENEZES, L.C.; PAES, A.; FINATTO, M.J.B. Investigação de técnica de aumento de dados por substituição lexical para apoio à simplificação textual automática. XV CELSUL – Simpósio, 2022

REFERÊNCIAS

AGGARWAL, Charu C. **Machine learning for text**. [S. l.]: Springer, 2018. v. 848.

ALMEIDA, Daniel Machado de; ALUÍSIO, Sandra Maria. Text readability analysis with natural language processing tools: The adaptation of coh-metrix metrics for Portuguese. In: IEEE. 2009 Seventh Brazilian Symposium in Information and Human Language Technology. [S. l.: s. n.], 2009. p. 53–62.

ALUISIO, Sandra et al. Readability assessment for text simplification. In: PROCEEDINGS of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. [S. l.: s. n.], 2010. p. 1–9.

ALUÍSIO, Sandra M et al. Towards brazilian portuguese automatic text simplification systems. In: PROCEEDINGS of the eighth ACM symposium on Document engineering. [S. l.: s. n.], 2008. p. 240–248.

BACCOURI, Nidhal. **Deep Translator**. [S. l.: s. n.], 2020. Disponível em: <https://deep-translator.readthedocs.io/en/latest/README.html>.

BAYER, Markus; KAUFHOLD, Marc-André; REUTER, Christian. A survey on data augmentation for text classification. **ACM Computing Surveys**, ACM New York, NY, 2021.

BENTZ, Christian et al. Complexity trade-Offs and equi-complexity in natural languages: A meta-analysis. **Linguistics Vanguard**, De Gruyter Mouton, 2022.

BICK, Eckhard. **The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework**. [S. l.]: Aarhus Universitetsforlag, 2000.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S. l.]: "O'Reilly Media, Inc.", 2009.

BRASIL, MEC/SECAD. **Coleção Literatura para Todos 1**. [S. l.: s. n.], 2006.

BRILL, Eric. Processing natural language without natural language processing. In: SPRINGER. INTERNATIONAL Conference on Intelligent Text Processing and Computational Linguistics. [S. l.: s. n.], 2003. p. 360–369.

BROWN, James Dean. An EFL readability index. **University of Hawai'i Working Papers in English as a Second Language 15 (2)**, 1997.

CARRELL, Patricia L. Readability in ESL. University of Hawaii National Foreign Language Resource Center, 1987.

CAYLOR, John S et al. Methodologies for Determining Reading Requirements of Military Occupational Specialties. ERIC, 1973.

CHAQUET-ULLDEMOLINS, Jacobo et al. On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders. **Applied Sciences**, MDPI, v. 12, n. 8, p. 3856, 2022.

CHARNIAK, Eugene. A maximum-entropy-inspired parser. In: 1ST Meeting of the North American Chapter of the Association for Computational Linguistics. [S. l.: s. n.], 2000.

CNDE. **Balanço do Plano Nacional de Educação 2021**. [S. l.: s. n.], 2021. Disponível em: <https://media.campanha.org.br/acervo/documentos/%20BALANCO_PNE_2021.pdf>.

CROSSLEY, Scott et al. A large-scaled corpus for assessing text readability. **Behavior Research Methods**, Springer, p. 1–17, 2022.

DALE, Edgar; CHALL, Jeanne S. A formula for predicting readability: Instructions. **Educational research bulletin**, JSTOR, p. 37–54, 1948.

DAVIES, Mark. **O corpus do português**. [S. l.: s. n.], 2017. Disponível em: <http://www.corpusdoportugues.org/>.

DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: BURSTEIN, Jill; DORAN, Christy; SOLORIO, Thamar (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S. l.]: Association for Computational Linguistics, 2019. p. 4171–4186. DOI: 10.18653/v1/n19-1423. Disponível em: <https://doi.org/10.18653/v1/n19-1423>.

DIAS-DA-SILVA, Bento Carlos; MORAES, Helio Roberto de. A construção de um thesaurus eletrônico para o português do Brasil. **ALFA: Revista de Linguística**, Universidade Estadual Paulista (UNESP), 2003.

EDOUARD, Grave et al. Learning Word Vectors for 157 Languages. In: PROCEEDINGS of the Eleventh International Conference on Language Resources and Evaluation. [S. l.]: European Language Resources Association, 2018. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html>.

EISENSTEIN, Jacob. **Natural language processing**. [S. l.]: MIT press, 2018.

FACEBOOK. **Learning Word Vectors for 157 Languages**. [S. l.: s. n.], 2018. Disponível em: <https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.pt.300.bin.gz>.

FENG, Lijun; ELHADAD, Noémie; HUENERFAUTH, Matt. Cognitively motivated features for readability assessment. In: PROCEEDINGS of the 12th Conference of the European Chapter of the ACL (EACL 2009). [S. l.: s. n.], 2009. p. 229–237.

FENG, Steven Y.; GANGAL, Varun et al. A Survey of Data Augmentation Approaches for NLP. In: FINDINGS of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2021. ACL/IJCNLP. (Findings of ACL), p. 968–988. DOI: 10.18653/v1/2021.findings-acl.84. Disponível em: <https://doi.org/10.18653/v1/2021.findings-acl.84>.

FERREIRA, Taynan Maier; COSTA, Anna Helena Reali. DeepBT and NLP Data Augmentation Techniques: A New Proposal and a Comprehensive Study. In: PROCEEDINGS of the 9th Brazilian Conference in Intelligent Systems. [S. l.]: Springer, 2020. v. 12319. (Lecture Notes in Computer Science), p. 435–449. DOI: 10.1007/978-3-030-61377-8_30. Disponível em: <https://doi.org/10.1007/978-3-030-61377-8%5C_30>.

FINATTO, Maria José Bocorny. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. **Estudos Linguísticos (São Paulo. 1978)**, v. 49, n. 1, p. 72–96, 2020.

FINATTO, Maria José Bocorny; PARAGUASSU, Liana Braga. Acessibilidade textual e terminológica. EDUFU, 2022.

FINATTO, Maria José Bocorny; TCACENCO, Lucas Meireles. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm: revista do Centro Interdepartamental de Tradução e Terminologia. São Paulo, SP. Vol. 37, n. 1 (jan. 2021), p. 30-63**, 2021.

FLESCH, Rudolph. A new readability yardstick. **Journal of applied psychology**, American Psychological Association, v. 32, n. 3, p. 221, 1948.

FRAZIER, Lyn. Syntactic complexity. **Natural language parsing: Psychological, computational, and theoretical perspectives**, p. 129–189, 1985.

FREITAS, Cláudia. **Linguística computacional**. [S. l.]: Parábola Editorial, 2022.

GAZZOLA, Murilo Gleyson; LEAL, SE; ALUISIO, Sandra Maria. Predição da complexidade textual de recursos educacionais abertos em português. In: SYMPOSIUM in Information and Human Language Technology - STIL. [S. l.]: SBS, 2019.

GOLDBERG, Yoav. A primer on neural network models for natural language processing. **Journal of Artificial Intelligence Research**, v. 57, p. 345–420, 2016.

GOOGLE. **Google Tradutor**. [S. l.: s. n.], 2023. Disponível em: <https://translate.google.com/>.

GRAY, William Scott; LEARY, Bernice Elizabeth. What makes a book readable. Univ. Chicago Press, 1935.

HARTMANN, Nathan; FONSECA, Erick et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: PROCEEDINGS of the 11th Brazilian Symposium in Information and Human Language Technology. [S. l.: s. n.], 2017.

HARTMANN, Nathan Siegle; ALUÍSIO, Sandra Maria. Adaptaçao lexical automática em textos informativos do português brasileiro para o ensino fundamental. **Linguamática**, v. 12, n. 2, p. 3–27, 2020.

HEATON, Jeff. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. **Genet. Program. Evolvable Mach.**, Springer, v. 19, n. 1-2, p. 305–307, 2018. DOI: 10.1007/s10710-017-9314-z. Disponível em: <https://doi.org/10.1007/s10710-017-9314-z>.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. Disponível em: <https://doi.org/10.1162/neco.1997.9.8.1735>.

HONNIBAL, Matthew; MONTANI, Ines. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. [S. l.], 2017.

IMPERIAL, Joseph Marvin. BERT Embeddings for Automatic Readability Assessment. In: PROCEEDINGS of the International Conference on Recent Advances in Natural Language Processing. [S. l.]: INCOMA Ltd., 2021. p. 611–618. Disponível em: <https://aclanthology.org/2021.ranlp-1.69>.

INC, Facebook. **fastText Library for efficient text classification and representation learning**. [S. l.: s. n.], 2022. Disponível em: <https://fasttext.cc/>.

JURASFKY, Daniel; MARTIN, James H. **An introduction to natural language processing, computational linguistics, and speech recognition**. [S. l.]: Pearson Education, Inc, 2020.

KINCAID, J Peter; FISHBURNE JR, Robert P et al. **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. [S. l.], 1975.

KINCAID, JP; YASUTAKE, JY; GEISELHART, R. **Use of the Automated Readability Index to assess comprehensibility of Air Force technical orders**. [S. l.], 1967.

KLARE, George Roger et al. Measurement of readability. Iowa State University Press, 1963.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. Ibict, 1996.

LANDAUER, Thomas K et al. How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In: PROCEEDINGS of the 19th annual meeting of the Cognitive Science Society. [S. l.: s. n.], 1997. p. 412–417.

LEAL, SE. **Nilc Metrix**. [S. l.: s. n.], 2022. Disponível em: <https://github.com/nilc-nlp/nilcmetrix>.

LEAL, SE. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. 2019. Tese (Doutorado) – Universidade de São Paulo.

LEAL, SE. **Simpligo Ranking**. [S. l.: s. n.], 2020. Disponível em: <http://fw.nilc.icmc.usp.br:23380/simpligo-ranking>.

LEAL, SE; DURAN, Magali Sanches; ALUÍSIO, Sandra. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In: PROCEEDINGS of the 27th International Conference on Computational Linguistics. [S. l.: s. n.], 2018. p. 401–413.

LEAL, SE; LUKASOVA, Katerina et al. RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation**, Springer, v. 56, n. 4, p. 1333–1372, 2022.

LEAL, SE; MAGALHAES, Vanessa Maia Aguiar de et al. Avaliação Automática da Complexidade de Sentenças do Português Brasileiro para o Domínio Rural. In: IN: SYMPOSIUM IN INFORMATION, HUMAN LANGUAGE TECHNOLOGY e COLLOCATES.

LEAL, SE; SANCHES DURAN, Magali et al. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **arXiv e-prints**, arxiv–2201, 2021.

LEAL, SE; SCARTON, Carolina et al. **NILC-Metrix**. [S. l.: s. n.], 2022. Disponível em: <http://fw.nilc.icmc.usp.br:23380/nilcmetrix>.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LEE, Bruce W.; JANG, Yoo Sung; LEE, Jason Hyung-Jong. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In: PROCEEDINGS of the 2021 Conference on Empirical Methods in Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2021. p. 10669–10686. DOI: 10.18653/v1/2021.emnlp-main.834. Disponível em: <https://doi.org/10.18653/v1/2021.emnlp-main.834>.

LEE, Denny; HEINTZ, Brenner. **Productionizing Machine Learning with Delta Lake**. [S. l.: s. n.], 2019. Disponível em: <https://databricks.com/blog/2019/08/14/productionizing-machine-learning-with-delta-lake.html>.

LINGUÍSTICA COMPUTACIONAL, NILC - Núcleo Interinstitucional de. **Repositório de Word Embeddings do NILC**. [S. l.: s. n.], 2017. Disponível em: <http://nilc.icmc.usp.br/embeddings>.

LIVELY, Bertha A; PRESSEY, Sidney L. A method for measuring the vocabulary burden of textbooks. **Educational administration and supervision**, v. 9, n. 7, p. 389–398, 1923.

LONGPRE, Shayne; WANG, Yu; DUBOIS, Christopher. How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers? **arXiv preprint arXiv:2010.01764**, 2020.

MA, Chuchu. **Readability Assessment with Pre-Trained Transformer Models: An Investigation with Neural Linguistic Features**. [S. l.: s. n.], 2022.

MARTIN, James H. **Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition**. [S. l.]: Pearson/Prentice Hall, 2009.

MARTINC, Matej; POLLAK, Senja; ROBNIK-ŠIKONJA, Marko. Supervised and Unsupervised Neural Approaches to Text Readability. **Computational Linguistics**, v. 47, n. 1, p. 141–179, abr. 2021. ISSN 0891-2017. DOI: 10.1162/coli_a_00398. eprint: https://direct.mit.edu/coli/article-pdf/47/1/141/1911429/coli_a_00398.pdf. Disponível em: <https://doi.org/10.1162/coli%5C_a%5C_00398>.

MARTINS, Teresa BF et al. **Readability formulas applied to textbooks in brazilian portuguese**. [S. l.]: Icmsc-Usp, 1996.

MCNAMARA DS., Max M. Louwerse; GRAESSER, Arthur C. Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. In: COGNITIVE Science and Educational Practice group at the University of Memphis. [S. l.: s. n.], 2002. Disponível em: <http://csep.psyc.memphis.edu/mcnamara/pdf/IESproposal.pdf>.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MILBURN, BA. **Curious Cases: A Collection of American and English Decisions, Selected for Their Readability**. [S. l.]: Michie Company, 1902.

MUNIZ, Marcelo Caetano Martins. **A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto Unitex-PB**. 2004. Tese (Doutorado) – Universidade de São Paulo.

NILC. **Repositório de Word Embeddings do NILC**. [S. l.: s. n.], 2017. Disponível em: <http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip_s300.zip>.

PANDAS. **API Reference: DataFrame**. [S. l.: s. n.], 2022. Disponível em: <https://pandas.pydata.org/docs/reference/frame.html>.

PASQUALINI, Bianca Franco. Corpop: um corpus de referência do português popular escrito do Brasil, 2018.

PEI, Mario; GAYNOR, Frank. **Dictionary of linguistics**. [S. l.]: Rowman & Littlefield, 1954.

PETERS, Matthew et al. Deep Contextualized Word Representations. In: PROCEEDINGS of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2018. 1 (Long Papers), p. 2227–2237.

PONOMARENKO, Gabriel Luciano. **Índices para cálculo de Leiturabilidade**. [S. l.]: Universidade Federal do Rio Grande do Sul, 2018. Disponível em: <http://www.ufrgs.br/%20textecc/acessibilidadett/files/Indices-de-Leiturabilidade.pdf>.

RADFORD, Alec et al. Improving language understanding by generative pre-training. OpenAI, 2018.

ŘEHŮŘEK, Radim. **Gensim Topic modelling for humans**. [S. l.: s. n.], 2022. Disponível em: <https://radimrehurek.com/gensim/index.html>.

RODRIGUES, Erica dos Santos; FREITAS, Cláudia; QUENTAL, Violeta. Análise de inteligibilidade textual por meio de ferramentas de processamento automático do português: avaliação da Coleção Literatura para Todos. **Letras de Hoje**, v. 48, n. 1, p. 91–99, abr. 2013. Disponível em: <https://revistaseletronicas.pucrs.br/index.php/fale/article/view/12048>.

ŞAHIN, Gözde Gül. To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP. **Computational Linguistics**, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA, v. 48, n. 1, p. 5–42, 2022.

SANTOS, Adriana Maximino dos. Leiturabilidade: É possível medi-la em livros infanto-juvenis? **Congresso Internacional de Leitura e Literatura Infantil e Juvenil**, Editora PUC-RS, 2010.

SARDINHA, Antonio Paulo Berber. **Corpus brasileiro: uma coletânea online de um bilhão de palavras do português brasileiro contemporâneo**. [S. l.: s. n.], 2010. Disponível em: <https://bv.fapesp.br/pt/auxilios/28549/corpus-brasileiro-uma-coletanea-online-de-um-bilhao-de-palavras-do-portugues-brasileiro-contemporane/>.

SCARTON, Carolina; GASPERIN, Caroline; ALUISIO, Sandra. Revisiting the readability assessment of texts in portuguese. In: SPRINGER. IBERO-AMERICAN Conference on Artificial Intelligence. [S. l.: s. n.], 2010. p. 306–315.

SCARTON, Carolina; OLIVEIRA, Matheus et al. SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In: PROCEEDINGS of the NAACL HLT 2010 Demonstration Session. [S. l.: s. n.], 2010. p. 41–44.

SCARTON, Carolina Evaristo; ALUÍSIO, Sandra Maria. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.

SCHWARM, Sarah E; OSTENDORF, Mari. Reading level assessment using support vector machines and statistical language models. In: PROCEEDINGS of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05). [S. l.: s. n.], 2005. p. 523–530.

SCIKIT-LEARN. **Modules: Logistic Regression**. [S. l.: s. n.], 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>.

SCIKIT-LEARN. **Modules: StandardScaler**. [S. l.: s. n.], 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.

SCIKIT-LEARN. **Modules: Support Vector Machines**. [S. l.: s. n.], 2021. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>.

SHI, Haoyue; LIVESCU, Karen; GIMPEL, Kevin. Substructure substitution: Structured data augmentation for NLP. **arXiv preprint arXiv:2101.00411**, 2021.

SHORTEN, Connor; KHOSHGOFTAAR, Taghi M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019.

SHUKLA, Anupriya; PANDEY, Hari Mohan; MEHROTRA, Deepti. Comparative review of selection techniques in genetic algorithm. In: IEEE. 2015 international conference on futuristic trends on computational analysis and knowledge management (ABLAZE). [S. l.: s. n.], 2015. p. 515–519.

SI, Luo; CALLAN, Jamie. A statistical model for scientific readability. In: PROCEEDINGS of the tenth international conference on Information and knowledge management. [S. l.: s. n.], 2001. p. 574–576.

SMITH, Edgar A; KINCAID, J Peter. Derivation and validation of the automated readability index for use with technical materials. **Human factors**, SAGE Publications Sage CA: Los Angeles, CA, v. 12, n. 5, p. 457–564, 1970.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9TH Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear). [S. l.: s. n.], 2020.

STOLCKE, Andreas. SRILM-an extensible language modeling toolkit. In: 7TH International Conference on Spoken Language Processing. [S. l.]: ISCA, 2002. Disponível em: <http://www.isca-speech.org/archive/icslp%5C_2002/i02%5C_0901.html>.

TAYLOR, Wilson L. "Cloze procedure": A new tool for measuring readability. **Journalism quarterly**, SAGE Publications Sage CA: Los Angeles, CA, v. 30, n. 4, p. 415–433, 1953.

THOMAS, Calvin et al. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In: IEEE. IEEE International Conference Mechatronics and Automation, 2005. [S. l.: s. n.], 2005. v. 3, p. 1569–1574.

THORNDIKE, EL. The Teacher's Word Book. **Columbia University, New York**, 1921.

UNIVERSITY, Syddansk. **Flat structure**. [S. l.: s. n.], 2022. Disponível em: <https://visl.sdu.dk/visl/pt/parsing/automatic/parse.php>.

USP, NILC. **TeP 2.0**. [S. l.: s. n.], 2022. Disponível em: <http://www.nilc.icmc.usp.br/tep2/ajuda.htm>.

USP, São Carlos. **PortiLexicon-UD a lexicon for Brazilian Portuguese according to Universal**. [S. l.: s. n.], 2022. Disponível em: <https://portilexicon.icmc.usp.br/>.

USP - SÃO CARLOS, FAPESP e IBM. **POeTiSA: POrtuguese processing - Towards Syntactic Analysis and parsing**. [S. l.: s. n.], 2022. Disponível em: <https://sites.google.com/icmc.usp.br/poetisa>.

VAJJALA, Sowmya; LUČIĆ, Ivana. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In: PROCEEDINGS of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018. p. 297–304. DOI: 10.18653/v1/W18-0535. Disponível em: <https://aclanthology.org/W18-0535>.

VANROSSUM, Guido; DRAKE, Fred L. **The python language reference**. [S. l.]: Python Software Foundation Amsterdam, Netherlands, 2010.

VASWANI, Ashish et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

WAGNER FILHO, Jorge A; WILKENS, Rodrigo; IDIART, Marco et al. The brwac corpus: A new open resource for brazilian portuguese. In: PROCEEDINGS of the eleventh international conference on language resources and evaluation (LREC 2018). [S. l.]: European Language Resources Association (ELRA), 2018. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/599.html>.

WAGNER FILHO, Jorge Alberto; WILKENS, Rodrigo; VILLAVICENCIO, Aline. Automatic construction of large readability corpora. In: PROCEEDINGS of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC). [S. l.: s. n.], 2016. p. 164–173.

WATANABE, Willian Massami et al. Facilita: reading assistance for low-literacy readers. In: PROCEEDINGS of the 27th ACM international conference on Design of communication. [S. l.: s. n.], 2009. p. 29–36.

WILKENS, Rodrigo; VECCHIA, Alessandro Dalla et al. Size does not matter. Frequency does. A study of features for measuring lexical complexity. In: SPRINGER. IBERO-AMERICAN conference on artificial intelligence. [S. l.: s. n.], 2014. p. 129–140.

WILKENS, Rodrigo; ZILIO, Leonardo et al. Crawling by readability level. In: SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [S. l.: s. n.], 2016. p. 306–318.

WOLF, Thomas et al. **Transformers: State-of-the-Art Natural Language Processing**. [S. l.]: Association for Computational Linguistics, 2020. p. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. Disponível em: <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.

WONG, Tzu-Tsung. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. **Pattern Recognition**, Elsevier, v. 48, n. 9, p. 2839–2846, 2015.

XIA, Menglin; KOCHMAR, Ekaterina; BRISCOE, Ted. Text Readability Assessment for Second Language Learners. **CoRR**, abs/1906.07580, 2019. arXiv: 1906.07580. Disponível em: <http://arxiv.org/abs/1906.07580>.

YNGVE, Victor H. A model and an hypothesis for language structure. **Proceedings of the American philosophical society**, JSTOR, v. 104, n. 5, p. 444–466, 1960.

YU, Adams Wei et al. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. **CoRR**, abs/1804.09541, 2018. arXiv: 1804.09541. Disponível em: <http://arxiv.org/abs/1804.09541>.