UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
IEG-INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VICTOR IVAN SILVA SILVEIRA

CLASSIFICAÇÃO DE LINGUAGEM SIMPLES: UMA ABORDAGEM BASEADA EM LEITURABILIDADE E LEGIBILIDADE

SANTARÉM-PA
2024

VICTOR IVAN SILVA SILVEIRA

CLASSIFICAÇÃO DE LINGUAGEM SIMPLES: UMA ABORDAGEM BASEADA EM LEITURABILIDADE E LEGIBILIDADE

Trabalho de Conclusão de Curso apresentado ao Programa de Computação, para obtenção do grau de Bacharel em Ciências da Computação; Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências.
Orientador: Prof. Dr. Fábio Manoel França Lobato

SANTARÉM-PA
2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/Ufopa

S587c Silveira, Victor Ivan Silva
Classificação de linguagem simples: uma abordagem baseada em leiturabilidade e legibilidade./ Victor Ivan Silva Silveira. – Santarém, 2024.
20 p.: il.
Inclui bibliografias.

Orientador: Fábio Manoel França Lobato.
Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Bacharelado em Ciência da Computação.

1. Linguagem simples. 2. Ciência de dados para governo. 3. Processamento de linguagem natural. I. Lobato, Fábio Manoel França, *orient*. II. Título.

CDD: 23 ed. 004

Bibliotecária - Documentalista: Renata Ferreira – CRB/2 1440

FORMULÁRIO DE AVALIAÇÃO DE TCC

Identificação:

Título do Trabalho:	Classificação de Linguagem Simples: uma abordagem baseada em Leiturabilidade e Legibilidade
Aluno (a): Victor Ivan Silva Silveira	
Orientador (a): Fábio Manoel França Lobato	

Avaliação:

Examinador (a) 1: Fábio Manoel França Lobato	**Nota:** 9,75
Assinatura:	

Examinador (a) 2: Liviane Ponte Rego	**Nota:** 9,6
Assinatura:	

Examinador (a) 3: Barbara Adriana Pires Barata	**Nota:** 9,9
Assinatura:	

Parecer:

A banca indicou que será necessário adicionar os elementos pré e pós-textuais

Resumo da Avaliação:

	Aceitação incondicional
X	Aceitação condicionada a modificações (especificar no verso)
	Recusado

Nota Final: 9,75

Santarém-PA, 26 de setembro de 2024.

gov.br Documento assinado digitalmente
FABIO MANOEL FRANCA LOBATO
Data: 26/09/2024 13:01:00-0300
Verifique em https://validar.iti.gov.br

_____

Presidente da Banca Examinadora

RESUMO

São inegáveis os esforços das agências de governo em adotar a linguagem simples. Não obstante, ainda é evidente a necessidade de ampliar esses esforços para outras esferas, uma vez que a linguagem simples é fundamental para permitir o entendimento claro da informação, promovendo a inclusão de cidadãos que não possuem o letramento suficiente. Na literatura, há diversos trabalhos dedicados à mensuração da linguagem simples e da complexidade textual, alguns utilizam métodos automáticos utilizando aprendizado de máquina, os quais carecem de explicabilidade sobre como é possível melhorar no texto; outros utilizam métodos semi-automáticos baseados na avaliação humana, o que impede a adoção do método para grandes conjuntos de dados. No presente trabalho, investigou-se o uso de métodos analíticos para classificação de Linguagem Simples à luz da Leiturabilidade e Legibilidade. Os resultados obtidos permitem concluir que é possível utilizar medidas de Leiturabilidade e Legibilidade para classificar a Linguagem Simples. O trabalho contribui para o estado da arte por meio do estudo de seis métricas de complexidade textual para classificação de Linguagem Simples. Para o estado da prática, o trabalho contribui com insumos para a construção de sistemas de classificação de Linguagem Simples, indicando aspectos de melhoria ao usuário final.

Palavras-Chave: Linguagem Simples, Ciência de Dados para Governo, Processamento de Linguagem Natural, Leiturabilidade, Legibilidade.

ABSTRACT

The efforts of government agencies to adopt Plain Language (PL) are undeniable. However, there is still a clear need to extend these efforts to other spheres since PL is fundamental to enabling a clear understanding of information and promoting the inclusion of citizens who are not sufficiently literate. In the literature, there are several studies dedicated to measuring PL and textual complexity, some of which use automatic methods using machine learning, which lack explanations of how it is possible to improve the text; others use semi-automatic methods based on human evaluation, which impairs its application on large data sets. We investigated the use of analytical methods for PL classification in the light of Readability and Legibility. The results show that using Readability and Legibility measures to classify Simple Language is possible. The work contributes to the state of the art by studying six textual complexity metrics for classifying Plain Language. For the state of the practice, the work contributes with inputs for developing and deploying Plain Language classification systems, indicating aspects of improvement for the end user.

Keywords: Plain Language, Data Science for Government, Natural Language Processing, Readability, Legibility.

SUMÁRIO

Classificação de Linguagem Simples: uma abordagem baseada em Leiturabilidade e Legibilidade

Victor I. S. Silveira[1], **Pedro H. C. Menezes**[1], **Marcelino S. Silva**[1], **Fabrício A. Carmo**[2], **Fábio M. F. Lobato**[1]

[1]Universidade Federal do Oeste do Pará (UFOPA)
Santarém – Pará – Brasil

[2]Universidade Estadual do Maranhão (UEMA)
São Luís - Maranhão - Brasil,

fabio.lobato@ufopa.edu.br

Abstract. *The efforts of government agencies to adopt Plain Language (PL) are undeniable. However, there is still a clear need to extend these efforts to other spheres since PL is fundamental to enabling a clear understanding of information and promoting the inclusion of citizens who are not sufficiently literate. In the literature, there are several studies dedicated to measuring PL and textual complexity, some of which use automatic methods using machine learning, which lack explanations of how it is possible to improve the text; others use semi-automatic methods based on human evaluation, which impairs its application on large data sets. We investigated the use of analytical methods for PL classification in the light of Readability and Legibility. The results show that using Readability and Legibility measures to classify Simple Language is possible. The work contributes to the state of the art by studying six textual complexity metrics for classifying Plain Language. For the state of the practice, the work contributes with inputs for developing and deploying Plain Language classification systems, indicating aspects of improvement for the end user.*

Resumo. *São inegáveis os esforços das agências de governo em adotar a linguagem simples. Não obstante, ainda é evidente a necessidade de ampliar esses esforços para outras esferas, uma vez que a linguagem simples é fundamental para permitir o entendimento claro da informação, promovendo a inclusão de cidadãos que não possuem o letramento suficiente. Na literatura, há diversos trabalhos dedicados à mensuração da linguagem simples e da complexidade textual, alguns utilizam métodos automáticos utilizando aprendizado de máquina, os quais carecem de explicabilidade sobre como é possível melhorar no texto; outros utilizam métodos semi-automáticos baseados na avaliação humana, o que impede a adoção do método para grandes conjuntos de dados. No presente trabalho, investigou-se o uso de métodos analíticos para classificação de Linguagem Simples à luz da Leiturabilidade e Legibilidade. Os resultados obtidos permitem concluir que é possível utilizar medidas de Leiturabilidade e Legibilidade para classificar a Linguagem Simples. O trabalho contribui para o estado da arte por meio do estudo de seis métricas de complexidade textual para classificação de Linguagem Simples. Para o estado da prática, o trabalho contribui com insumos para a construção de sistemas de classificação de Linguagem Simples, indicando aspectos de melhoria ao usuário final.*

1. Introdução

É inegável a crescente cobrança às organizações, principalmente governamentais, em relação à competência de oferecer clareza sobre o seu funcionamento, desempenho e resultados [Fung et al. 2007]. No Brasil, leis como Lei de Acesso à Informação e Lei complementar de Transparência, foram criadas visto a crescente necessidade de disponibilização de informações públicas. É dever das organizações tornar acessível os dados em seus sítios institucionais. Entretanto, a transparência não apenas se resume em acesso aos dados, o conceito abrange aspectos que visam fornecer não só informações de acesso geral, mas também permitir o entendimento dos dados [Aló and Leite 2009]. Nesse sentido, a Linguagem Simples (LS) surge como abordagem primordial a fim de promover a comunicação acessível e transparente [Cappelli et al. 2021].

A definição do termo Linguagem Simples, oriundo do inglês *Plain Language*, não é trivial. Apesar disso, grande parte dos estudiosos indicam que a LS visa revelar o conhecimento de forma clara e simplificada, tendo como seus principais objetivos, a construção de textos compreensíveis e utilizáveis [Petelin 2010]. Desse modo, LS consiste em um conjunto de diferentes abordagens que visam reduzir a complexidade e o uso de jargões, por meio de iniciativas de simplificação de textos, visando torná-lo mais acessível a diversos públicos [Cappelli et al. 2023]. Diante disso, o uso da técnica de LS se torna fundamental, visto que permite inclusão social a partir de transmissão de informações de maneira acessível. Tal abordagem está presente nas mais diversas áreas gerais, como saúde [Hildenbrand et al. 2020, Lyu et al. 2023], educação [Hansen-Schirra and Maass 2020], jurídica [Maass 2020], comercial [Rashid and Rasheed 2024], dentre outros. Apesar da implementação da LS possa variar dependendo do público alvo, existem diretrizes comuns, como o uso de frases curtas e diretas, uso de vocabulário comum e a estruturação da informação de forma clara. Por meio disso, promove-se a inclusão de acessibilidade à informação, pois ao simplificar a linguagem, a leiturabilidade e legibilidade atingem seu ápice. Desse modo, guias de LS estão tornando-se comuns na literatura.

À luz do contexto apresentado, o objetivo desse trabalho é classificar de textos com base no grau de linguagem simples à luz da leiturabilidade e legibilidade. Foram utilizados métodos analíticos aliados às abordagens que utilizam algoritmos de *Machine Learning* (ML). Desse modo, foi empregado os índices de *Flesch*, *Gunning*, *Automated Readability*, *Coleman* e *Gulpease*, pois representam métodos analíticos já consolidados no estado da arte e prática, assim como, os algoritmos *Support Vector Machine* (SVM) e o *Stochastic Gradient Descendent* (SGD). Os resultados obtidos contribuem para o estado da arte trazendo à baila conceito de leiturabilidade conjugado com LS, provendo medidas de avaliação objetivas e já validadas em outras aplicações. Para o estado da prática, o trabalho provê um ferramental que pode automatizar a avaliação da LS, reduzindo a subjetividade inerente e aumentando a produtividade dos avaliadores.

O restante deste artigo está organizado como segue. Na Seção 2, apresentam-se conceitos pertinentes ao trabalho. Na Seção 3, descrevem-se trabalhos relacionados a esta pesquisa. Na Seção 4 descrevem-se os materiais e métodos utilizados. Na Seção 5 é apresentado os resultados e os *insights* oriundo das análises. Por fim, a Seção 6 apresenta as considerações finais e trabalhos futuros.

2. Linguagem simples, leiturabilidade e conceitos correlatos

[Nord 2018] pontua que Linguagem Simples possui duas concepções. De acordo com o primeiro conceito, a LS pode ser caracterizada como um ideal estilístico de textos, representado por uma sintaxe simples, predominância de verbos na voz ativa, poucas ou ausência de palavras complexas, ou especializadas, dentre outros fatores. Por outro lado, a segunda compreensão enfatiza que a LS está associada ao conceito de adaptação ao destinatário, na qual o efeito textual é determinado pela mentalidade do escritor que elabora o texto visando às necessidades de seu público.

[Kamandhari 2020] afirma que leiturabilidade se concentra na ideia da compreensão textual e a sua medição é determinada pelo tempo gasto na leitura de um texto sem qualquer dificuldade. Além disso, aponta que tal pode ser definida a partir de dois fatores distintos, sendo: os conceitos extraídos de fatores relacionados com o texto e o impacto emocional devido ao *design* do texto. De acordo com [Bailin and Grafstein 2016], leiturabilidade se preocupa com a comunicação eficaz de ideias e informações, sendo definida pela variedade de fatores linguísticos, abrangendo propriedades sintáticas, semânticas, morfológicas e textuais. Ademais, afirma-se que índices de leiturabilidade são utilizados para classificar a dificuldade de leitura dos textos. Conforme [Srisunakrua and Chumworatayee 2019], leiturabilidade enfatiza a garantia que o material de leitura corresponda à competência do leitor alvo. Sustenta, também, que um alto nível de leiturabilidade em um material expressa dificuldade na leitura, enquanto um material com baixo nível de leiturabilidade é considerado simples de ler.

Segundo [Tekfi 1987], legibilidade é definido como estudo da visibilidade e perceptibilidade da informação presente em um texto a fim de fornecer compreensibilidade ao leitor por meio da facilidade, velocidade e precisão no qual as informações são obtidas mediante aos elementos textuais presentes. Além de tais características, também é destacado o tamanho da fonte, espaçamento linear, comprimento das linhas em relação a caracteres, margem, entre outros. O trabalho de [Kamandhari 2020] apresenta alguns subtemas a fim de conceituar legibilidade, a saber: forma específica dos caracteres abrangendo a facilidade ou dificuldade de distinção de um caractere ou letra em relação ao formato, visão expandida de um caractere concentrando na sua constituição e relação sendo letra, símbolo, número ou pontuação, e envolvimento dos fatores tipográficos determinado pelo espaço em branco entre letras, estilo, tamanho e espessura da fonte. Em resumo, cada subtema justifica a importância e correlação dos elementos específicos para que seja possível o compreendimento textual a partir da extração de informação por meio de uma leitura simples devido à presença dos componentes anteriormente apresentados.

Importante pontuar que apesar de diferentes, os conceitos legibilidade e leiturabilidade frequentemente se referem a um mesmo fenômeno, o quão difícil é ler um texto - tendendo mais para a leiturabilidade. Mesmo assim, optamos por manter a terminologia utilizada originalmente, tal como a primeira métrica que é apresentada a seguir, o índice de legibilidade Flesch (*Flesch Reading Ease*), que é responsável por avaliar a compreensibilidade textual por meio de uma fórmula que incorpora métricas como contagem de palavras, sentenças e sílabas, representada pela Equação 1. Tal prática se torna possível com o surgimento e evolução da computação em termo de processamento.

$$\textit{Flesch Reading Ease} = 206,835 - \left(1,015 \times \frac{\text{palavras}}{\text{sentenças}}\right) - \left(84,6 \times \frac{\text{sílabas}}{\text{palavras}}\right) \quad (1)$$

O índice de nebulosidade de *Gunning* (*Gunning Fog Index*) tem a finalidade em estimar o grau de educação formal a partir de compreensão de texto sem apresentar qualquer dificuldade na leitura. Nesse índice faz o uso da métrica "palavras complexas", definida por *Gunning* que palavras as quais possuem mais de três sílabas devem ser classificadas como tal. A Eq. 2 demonstra a estrutura de tal índice.

$$\textit{Gunning Fog index} = 0,4 \times \left(\frac{\text{palavras}}{\text{sentenças}}\right) + 40 \times \left(\frac{\text{palavras complexas}}{\text{palavras}}\right) \quad (2)$$

O Índice de Legibilidade Automatizado (*Automated Readability Index* - ARI) possui o diferencial de não necessitar de contagem de sílabas, dada a complexidade dessa tarefa. Tal abordagem é destacada pela facilidade em quantificar a legibilidade, visto que contagem de palavras, caracteres e sentenças são consideradas tarefas simples. O índice está representado pela Eq. 3, a seguir.

$$\text{ARI} = -21,43 + 0,5 \times \left(\frac{\text{palavras}}{\text{sentenças}}\right) + 4,71 \times \left(\frac{\text{caracteres}}{\text{palavras}}\right) \quad (3)$$

O nível de instrução de *Flesch-Kincaid* (*Flesch-Kincaid grade level*) é caracterizada pela reformulação das equações 1, 2 e 3, em que é utilizada a escala zero a cem após a conversão. Devido às manipulações, tem-se a Eq. 4.

$$\text{Flesch-Kincaid} = 63,88 - 0,38424 \times (FleschReadindEase) - 20,7 \times \left(\frac{\text{sílabas}}{\text{palavras}}\right) \quad (4)$$

O Índice de *Coleman-Liau* (*Coleman-Liau Index*) foi criado com a finalidade de prover uma implementação simples, sendo o mesmo intuito de ARI (Eq. 3). A diferença desse método é caracterizado pela inversão da primeira razão da abordagem ARI (Eq. 3), resultando em sentenças/palavras. Tais alterações são representadas pela Eq. 5.

$$\text{Coleman-Liau} = -15,8 - 2,96 \times \left(\frac{\text{sentenças}}{\text{palavras}}\right) + 5,88 \times \left(\frac{\text{letras}}{\text{palavras}}\right) \quad (5)$$

O Índice de *Gulpease* foi criado com o objetivo de estimar a legibilidade de textos na língua italiana a partir da Eq. 6 a qual adota a escala de zero a cem e, além disso, o uso de tal método é viável devido a não utilização de contagem de sílaba, isto é, não há complexidade computacional para essa tarefa.

$$\text{Gulpease Index} = 89 + 300 \times \left(\frac{\text{sentenças}}{\text{palavras}}\right) - 10 \times \left(\frac{\text{letras}}{\text{palavras}}\right) \quad (6)$$

3. Trabalhos Relacionados

No trabalho de [Moreno et al. 2022], foi proposto um *software* chamado Análise de Legibilidade Textual (ALT[1]), cujo principal objetivo consiste em medir e avaliar os índices de legibilidade de textos dispostos na língua portuguesa. Para isso foram consideradas os seguintes abordagens: *Flesch reading ease*, *Gunning fog index*, *Automated readability index*, *Flesch-Kincaid grade level*, *Coleman-Liau index* e *Gulpease index*. Considerando que estes modelos foram originalmente implementados para idiomas diferentes do

[1] `https://https://legibilidade.com/`

português, os autores realizaram um processo de adaptação de seus parâmetros a partir da regressão linear de seus resultados, utilizando um conjunto de treinamento composto de textos da língua portuguesa traduzidos para o idioma padrão dos métodos (inglês e francês). Os resultados evidenciaram uma correlação acima de 97% entre os resultados das fórmulas aptadas e a sua respectiva original. Visando expandir a metodologia utilizada por [Moreno et al. 2022], o presente estudo foca em abranger conceitos relacionados não só à legibilidade, mas também à leiturabilidade.

[Rodrigues et al. 2023] apresentam uma proposta para a construção da ferramenta de automação do Índice Nacional de Avaliação de LS, a fim de avaliar textos presentes em portais públicos com base nas boas práticas definidas no Guia Nacional de LS. A metodologia de avaliação possui quatro etapas: a) Coleta de informações; b) Análise de evidências; c) Cálculo do índice; d) Apresentação dos resultados. Os artefatos produzidos incluem o protótipo do *software* e *dataset* com palavras classificadas a partir do grau de dificuldade. Os autores utilizaram metodologias de avaliação já consolidadas na literatura, juntamente com técnicas de ML, com foco a estimar o grau de legibilidade, complexibilidade/leiturabilidade de textos na língua portuguesa a fim de classificar a LS.

[Dressler et al. 2023] utilizaram de modelos computacionais para a classificação da complexidade do texto. Foram utilizados algoritmos de ML, aliado as técnicas de *Natural Language Processing* (NLP). Para a realização dos experimentos foi utilizado o *dataset* chamado "*Corpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro*" a fim de treinar os modelos de ML. Esse *corpus* consiste em um conjunto de textos didáticos rotulados segundo o nível de escolaridade. Comparando diversos modelos de aprendizado de máquina, os autores concluíram que os métodos baseados em otimização: *Multilayer Perceptron* (MLP) e *Stochastic Gradient Descendent* (SGD) apresentaram um desempenho superior. Entretanto, há a carência de comparação dos resultados com as metodologias de cálculo da complexidade já presentes na literatura. Em vista dessa limitação, o presente trabalho visa sanar essa lacuna por meio da investigação de métodos analíticos para classificar a complexidade textual.

[Moutinho and Picanço 2022] verificaram a compatibilidade de textos didáticos com o respectivo ano escolar (5º a 9º ano). Foi utilizado índice Flesch adaptado ao português para medir a complexidade dos textos presentes nos livros. Os autores utilizaram a ferramenta chamada Coh-Metrix-Port. Os textos utilizados na pesquisa são oriundos de livros didáticos presente em portais eletrônicos, como o E-Docente e os respectivos sites das editoras. Foram selecionados três textos a cada ano escolar, todos correspondendo ao mesmo tema e próximos em quantidade de palavras. Os achados demonstram uma disparidade entre o ano escolar e a complexidade dos textos, uma vez que, a metodologia de medição de leiturabilidade aponta que parte considerável dos textos são classificados como de ensino médio e ensino superior. O presente trabalho busca ampliar as metodologias para o cálculo da complexidade, assim como comparar com outras abordagens para a classificação de textos escolares à luz da LS.

A partir da análise dos trabalhos relacionados, ficou evidente que tanto técnicas de ML, quanto metodologias de cálculo de índice, apresentam relevância na literatura atual na esfera da LS. Ademais, os estudos revelaram que diversas abordagens de metodologias para o cálculo da Leiturabilidade e Legibilidade, são pertinentes para classificar textos com base na complexidade. Ademais, diversos estudos comparam abordagens de

ML, enquanto outros estudos, comparam metodologias baseadas em índice. No entanto, nota-se uma ecasses de trabalhos que associam ambas as formas de classificação da LS de textos em português. Desse modo, o presente estudo tem em vista preencher esta lacuna presente na literatura, com o foco em utilizar tanto abordagens de ML, quanto metodologias analíticas em vista da LS.

4. Materiais e Métodos

No presente trabalho, o método de pesquisa *Data Science Trajectories model* (DST) foi escolhido em razão de sua flexibilidade e liberdade de escolha para com as etapas do projeto, delimitando e otimizando os processos a serem seguidos. Este modelo foi proposto por [Martínez-Plumed et al. 2019], consistindo em uma expansão da metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) desenvolvido por [Wirth and Hipp 2000]. Seu diferencial é caracterizado por seu percurso ser definido pelo(a) pesquisador(a) a partir de processos mais abrangentes em relação ao CRISP-DM em consequência da inclusão de atividades mais exploratórias, como exploração de objetivos, exploração de fontes de dados e exploração de valor de dados.

Portanto, a utilização do DST é justificada por três motivos primordiais, sendo: (1) possibilidade dos cientistas de dados definirem o conjunto das atividades, explorando novas etapas que podem ser adicionadas ou removidas do seu fluxo de trabalho, que podem ser devidamente encapsuladas e documentadas bem como no CRISP-DM; (2) catálogo de trajetórias serve como referência, possibilitando a correspondência para novos projetos e evitando o ajuste forçado a um modelo de processo restritivo como o CRISP-DM; (3) as trajetórias podem ser mapeadas em detrimento ao plano do projeto, permitindo atribuição de prazos às transições de etapas e modelos espirais denotados por iterações explícitas na trajetória. No presente estudo foram utilizadas as seguintes etapas: Entendimento do negócio; Aquisição dos dados; Entendimento dos dados; Preparação dos dados; Modelagem; Exploração dos resultados; e Avaliação.

Entendimento do negócio: Esta etapa é considerada o ponto fundamental de um projeto de ciência de dados em razão de conceber a análise, compreensão e definição do escopo total do projeto [Schäfer et al. 2018]. Foram tomados passos como: definição da base do negócio, descrição dos problemas encontrados, objetivos a serem alcançados e foco do projeto. Neste sentido, os conceitos sobre LS, leiturabilidade e legibilidade foram explorados na literatura a fim de possibilitar a realização de tal pesquisa. O resultado de tal passo foi apresentado na Seção 2. Além disso, foi estabelecido as devidas metas a serem alcançadas ao longo do projeto, tais são: entender e aplicar os conceitos de LS na tarefa de classificação textual; aperfeiçoar os métodos de cálculo de legibilidade por meio de adaptações dos índices; e aliar métodos analíticos e modelos inteligentes dada a mesma atividade.

Aquisição dos dados: Os dados disponibilizados por [Murilo Gazzola 2019][2] foram utilizados nos experimentos. O conjunto de dados é composto por 2.076 documentos textuais representando diferentes níveis de escolaridade do sistema educacional brasileiro. Tais documentos já estão anotados/classificados de acordo com quatro estágios: ensino fundamental I; ensino fundamental II; ensino médio e ensino superior.

[2] `https://github.com/gazzola/corpus_readability_nlp_portuguese`

Entendimento dos dados: Essa fase é caracterizada pela familiarização para com os dados, identificando problemas em relação à qualidade dos dados. Também, é possível obter primeiros *insights* a fim de formar hipóteses para reconhecer informações ocultas [Nadali et al. 2011]. Com isso, foi realizado uma breve análise exploratória com a utilização da biblioteca Pandas e linguagem *Python*. Em relação às características do conjunto de dados, observa-se um alto grau de desbalanceamento entre as classes. No entanto, é importante ressaltar que essas classes não foram diretamente consideradas nos experimentos que visavam mensurar os índices de complexidade da linguagem. Em vez disso, elas foram utilizadas para verificar se as classes de complexidade estão diretamente relacionadas com os níveis de escolaridade.

Preparação dos dados: Nessa etapa, foram mapeadas duas variações do conjunto de dados considerando as técnicas de classificação de linguagem utilizadas. Para os métodos analíticos não foram necessários procedimentos de limpeza de dados, visto que elementos como pontuações, *stopwords* e símbolos especiais são relevantes. No entanto, para a análise utilizando os algoritmos de ML, etapas de pré-processamento são fundamentais. Nesse sentido, foram aplicados vários filtros de limpeza e configuração dos textos, como: remoção de *stopwords*, pontuações e caracteres especiais, configuração dos textos para *lowercase*.

Modelagem: A fase de modelagem foi mapeada em duas etapas: i) Extração das características textuais e ii) Classificação da Linguagem. A primeira envolve o processo de extração dos parâmetros necessários (*e.g.*, números de palavras, caracteres, sílabas) e a segunda contempla o processo de construção e aplicação dos modelos de classificação da linguagem (analíticos e de ML). As características foram obtidas por meio de Expressões Regulares e com recursos da biblioteca Python NLTK[3]. A Tabela 1 apresenta e descreve o conjunto de características extraídas dos documentos.

Tabela 1. Conjunto de caraterísticas textuais utilizadas.

Id	Característica	Descrição
1	Caracteres	Realiza a contagem de todos os caracteres de um documento
2	Letras	Realiza a contagem apenas das letras de um documento
3	Sílabas	Realiza a contagem de sílabas de um documento
4	Palavras	Realiza a contagem de palavras de um documento
5	Palavras complexas	Realiza a contagem de palavras complexas de um documento
6	Sentenças	Realiza a contagem das sentenças de um documento

Para o processo de identificação de uma palavra complexa foi determinado de acordo com o número de sílabas. Uma palavra é considerada complexa se possuir mais de duas sílabas [Moreno et al. 2022]. As sentenças são contabilizadas fazendo um *split* no texto, considerando ponto final, ponto de exclamação, ponto de interrogação ou ponto e vírgula. As demais características estão descritas na Tabela 1. Para a segunda etapa, foram selecionados os seguintes modelos analíticos para o cálculo do nível de legibilidade: *Flesch-Kincaid Index*; *Gunning Fog Index*; *Automated Readability Index*; e *Coleman-Liau Index*. O resultado é dado por meio de uma média simples de quatro índices. Os índices foram selecionados por já serem consolidados na literatura e por possuírem uma escala

[3]https://www.nltk.org/

intervalar para determinar a legibilidade de determinado texto. O nível complexidade é dado da seguinte forma: i) **Complexidade baixa**: se a média dos índices for inferior à 13 (< 13); **Complexidade média**: se a média dos índices for maior ou igual à $13 (>= 13)$ e menor que 17 (< 17); e **Complexidade alta:** se a média for maior que 17.

Além da utilização dos métodos analíticos já mencionados, foram utilizados dois modelos de classificação de dados: *Stochastic Gradient Descent* (SGD) e *Support Vector Machine* (SVM) visando o aprendizado das classes/níveis de linguagem determinadas pelos modelos analíticos. O SGD é uma técnica de otimização simples e eficiente para ajustar classificadores e regressores lineares, em que não se caracteriza como algoritmo de ML. Por outro lado, o SVM é um algoritmo que cria um hiperplano para diferenciação das classes e as classifica com apoio de vetores de suporte. Ambas técnicas, anteriormente citadas, são utilizadas em classificação textual e tarefas de NLP. Para o processo de vetorização dos documentos, utilizou-se a abordagem de Características Estatísticas Textuais (CET) [Almeida do Carmo et al. 2023]. Essa abordagem contempla o mapeamento de 10 características textuais, incluindo aquelas utilizadas neste trabalho (conforme apresentado na Tabela 1), para a formação do vetor. Dessa forma, incorporamos os traços característicos de cada classe de complexidade.

Exploração dos resultados: Consistiu em relacionar os resultados de ciência de dados com os objetivos de negócio. Por meio de análise estatística foi possível identificar dois *outliers* nos níveis de legibilidade textual, representados por valores máximos tão altos das complexidades média e alta. Nessa situação, fez-se necessário realizar uma investigação em que identificou-se manualmente um elevado número de sílabas, nos dois textos, que resultou no nível de legibilidade tão elevado (*outliers*). Além disso, notou-se que os valores da variância demonstram uma dispersão nos resultados obtidos, ou seja, há uma irregularidade nos valores alcançados por meio da aplicação da média final dos índices nos respectivos textos pertencentes as complexidades média e alta. **Avaliação:** A avaliação dos resultados se deu de forma quantitativa, utilizando de matrizes de confusão dos classificadores e a análise de sensibilidade das medidas de Leiturabilidade.

5. Resultados

Considerando as configurações experimentais, mencionadas na Seção 4, o cálculo dos índices foram computados e a média entre eles foi adotada para o processo de classificação do conjunto de dados de acordo com os índices de complexidade: baixa, média ou alta. A Tabela 2 apresentam os resultados dos modelos analíticos estudados.

Tabela 2. Configuração das classes de complexidade de acordo com os índices de Legibilidade aplicados.

Classe de Complexidade	Média dos índices					
	Quant. (%)	**Flesch Kincaid**	**Gunning**	**A.R**	**Coleman Liau**	**Média**
Baixa	800 (38,5%)	8.78	13.86	10.03	11.42	11.02
Média	1009 (48,6%)	12.81	17.76	14.46	14.13	14.79
Alta	267 (12,9%)	20.65	27.89	24.08	15.65	22.07

Com a aplicação dos índices de legibilidade notou-se uma nova configuração do conjunto de dados original. A Tabela 2 apresenta a distribuição das classes obtidas. Os

resultados mostram que a maioria dos documentos que compõe a conjunto de dados têm complexidade média (48,6%) e a minoria são os documentos com alta complexidade, cerca de 12%. Desse forma, não é possível indicar um paralelo com as classes de níveis de ensino. A Figura 1 ressalta essa diferenças entre os dois processos de anotações.

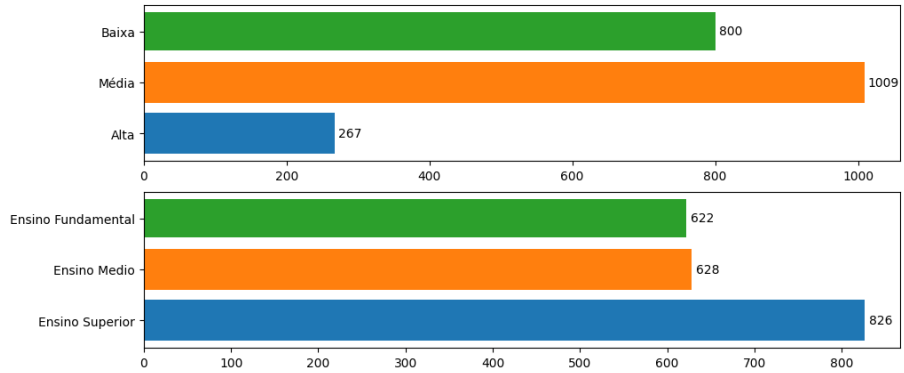


Figura 1. Comparação das distribuições de classes de complexidade e níveis de ensino no conjunto de dados educacional.

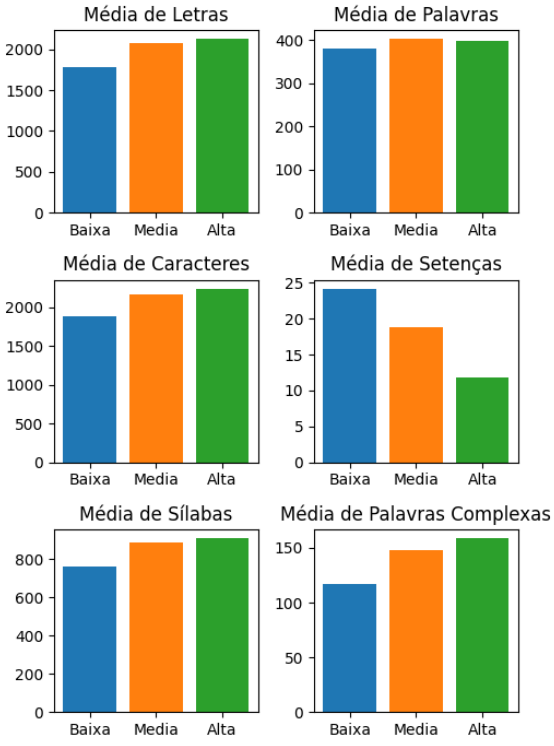


Figura 2. Características textuais extraídas do conjunto de dados educacional.

Em uma análise ao nível das características textuais (Tabela 2), observou-se que os documentos com complexidade baixa possuem médias inferiores para cinco das características adotadas. Isso é evidenciado com maior destaque na Média de Palavras Complexas. Tais evidências corroboram aos trabalhos correlatos apresentados, os quais apontam que textos considerados simples possuem sintaxe simples, com até duas sílabas. Por outro lado, documentos rotulados como complexos possuem alto índice de palavras complexas.

Outro ponto importante sobre aos textos considerados complexos, vide Figura 2, é que eles se caracterizam por possuírem sentenças de tamanhos inferiores as demais classes. Os resultados mostram que quanto maior é a complexidade, menor é o tamanho da sentença. Com a nova configuração do conjunto de dados, aplicaram-se procedimentos de classificação de dados utilizando os algoritmos SGD e SVM. O foco está na verificação do aprendizado dos padrões que caracterizam cada um dos tipos de complexidade. Os resultados para a acurácia do classificador foram de 81% e 85% para SGD e SVM, respectivamente. A Figura 3 apresenta a matriz de confusão com o aprendizado do modelo SVM. Ressalta-se que não foi aplicado procedimentos de balanceamento das classes.

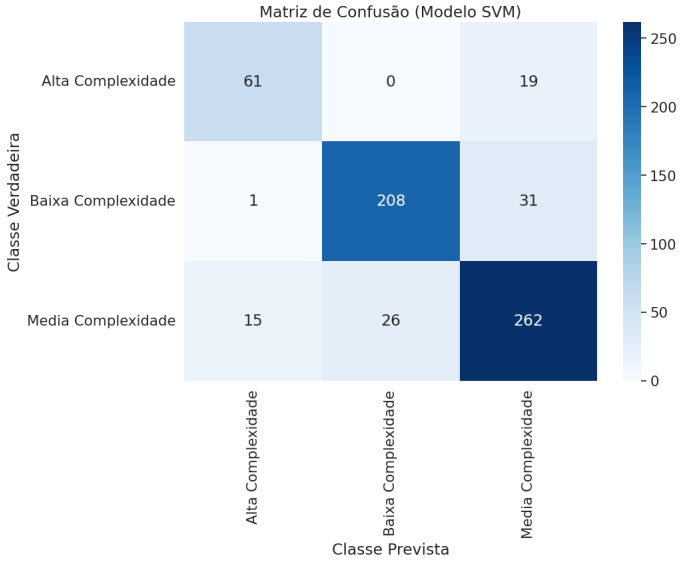


Figura 3. Resultado do modelo treinado com SVM.

Os resultados mostram que por meio das características textuais que mapeiam os níveis de complexidade de um documento é possível treinar modelos preditores para classificação da LS. Destaca-se que o processo de vetorização dos textos de entrada foi realizado através do CET, uma abordagem que considera esses traços característicos advindos da estrutura textual.

6. Considerações Finais

A LS é um aspecto fundamental para a democracia, possibilitando acesso à informação e inclusão dos cidadãos. Diversos trabalhos visam classificar textos consoante a linguagem simples, alguns utilizam métodos de aprendizado de máquina e outros usam métodos semiautomáticos. Pelo nosso melhor conhecimento, não há na literatura método que conjugue conceitos de Linguagem Simples e Leiturabilidade, além de permitir uma interpretabilidade do modelo, fornecendo ao usuário subsídios para a melhora do texto. Visando preencher esta lacuna, este trabalho apresentou o uso de medidas de legibilidade e leiturabilidade para classificação de linguagem simples, com o diferencial de não apenas automatizar a tarefa, mas fornecer informações que guiem a melhora do texto.

Para tal, foram avaliados seis métricas de legibilidade, a saber: *Flesch Reading Ease*, *Gunning Fog Index*, *Automated Readability Index*, *Flesch-Kincaid grade level*, *Coleman-Liau Index* e o *Gulpease Index*; e dois classificadores, o *Support Vector Machine* (SVM) e o *Stochastic Gradient Descendent* (SGD). Os resultados obtidos se mos-

traram consoantes com a literatura, com o SVM apresentando melhor desempenho para este domínio de aplicação. Ademais, a representação textual foi enriquecida, visto que o trabalho teve como diferencial a adoção de Características Estatísticas Textuais (CET).

Neste sentido, o trabalho contribui para o estado da arte por meio do estudo de seis métricas de complexidade textual para classificação de linguagem simples. Para o estado da prática, o trabalho contribui com insumos para a construção de sistemas de classificação de linguagem simples, indicando aspectos de melhoria ao usuário final. Além disso, destaca-se a construção do *dataset* formado com textos escolares classificados à luz da complexidade. Apesar dos esforços para uma ampla cobertura experimental, há espaços para melhora do estudo, com destaque para o balanceamento do *dataset* e adoção de outros conjuntos de dados. Vislumbra-se, também, a produtificação da ferramenta a fim de integrar um sistema que auxilie comunicadores na adequação de seus textos para linguagem simples, potencializando os impactos sociais do trabalho.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9, PIBITI - 100795/2024-2; e pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM- nº 045/2021.

Referências

Almeida do Carmo, F., Figueira da Silva Junior, J. L., Geraldeli Rossi, R., and França Lobato, F. M. (2023). Text representations for lyric-based identification of musical subgenres. *IEEE Latin America Transactions*, 21(6):737–744.

Aló, C. C. and Leite, J. d. P. (2009). Uma abordagem para transparência em processos organizacionais utilizando aspectos. *Rio de Janeiro*.

Bailin, A. and Grafstein, A. (2016). *Readability: Text and context*. Springer.

Cappelli, C., Nunes, V., and Oliveira, R. (2021). Transparência e transformaçao digital: O uso da técnica da linguagem simples. *Sociedade Brasileira de Computação*.

Cappelli, C., Oliveira, R., and Nunes, V. (2023). Linguagem simples como pilar da transparência. *Humanidades & Inovação*, 10(9):32–45.

Dressler, N., Souza, A. C., Costa, L. M., Souza, F. C., and Mantovani, R. (2023). Classificação de textos escolares com aprendizado de máquina. *Anais do Computer on the Beach*, 14:432–438.

Fung, A., Graham, M., and Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge University Press.

Hansen-Schirra, S. and Maass, C. (2020). Easy language, plain language, easy language plus: perspectives on comprehensibility and stigmatisation. *Easy language research: text and user perspectives*, 2:17.

Hildenbrand, G. M., Perrault, E. K., and Keller, P. E. (2020). Evaluating a health literacy communication training for medical students: Using plain language. *Journal of health communication*, 25(8):624–631.

Kamandhari, H. H. (2020). The definitions and the measurement of legibility and readability in instructional text design: an integrated literature review. *Information, Medium and Society*, 18(2):1.

Lyu, Q., Tan, J., Zapadka, M. E., Ponnatapura, J., Niu, C., Myers, K. J., Wang, G., and Whitlow, C. T. (2023). Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.

Maass, C. (2020). *Easy language–plain language–easy language plus: Balancing comprehensibility and acceptability*. Frank & Timme.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8):3048–3061.

Moreno, G. C. d. L., de Souza, M. P., Hein, N., and Hein, A. K. (2022). Alt: um software para análise de legibilidade de textos em língua portuguesa. *arXiv preprint*.

Moutinho, M. and Picanço, G. (2022). Índices de leiturabilidade e os textos didáticos: uma questão a ser discutida. *Língu@ Nostr@*, 10(2):124–147.

Murilo Gazzola, Sidney Evaldo Leal, S. M. A. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Nadali, A., Kakhky, E. N., and Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on crisp-dm methodology by a fuzzy expert system. In *2011 3rd International Conference on Electronics Computer Technology*.

Nord, A. (2018). Plain language and professional writing: A research overview.

Petelin, R. (2010). Considering plain language: issues and initiatives. *Corporate Communications: An International Journal*, 15(2):205–216.

Rashid, D. A. and Rasheed, D. R. (2024). Logistics service quality and product satisfaction in e-commerce. *SAGE Open*, 14(1):21582440231224250.

Rodrigues, A. P., Marques, G. M., Rodrigues, L. B., Mattos, P. A. A., Nunes, V. T., Cappelli, C., Oliveira, R., and de Moraes, R. M. (2023). Uma proposta de automação para o índice nacional de avaliação de linguagem simples em serviços públicos. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*. SBC.

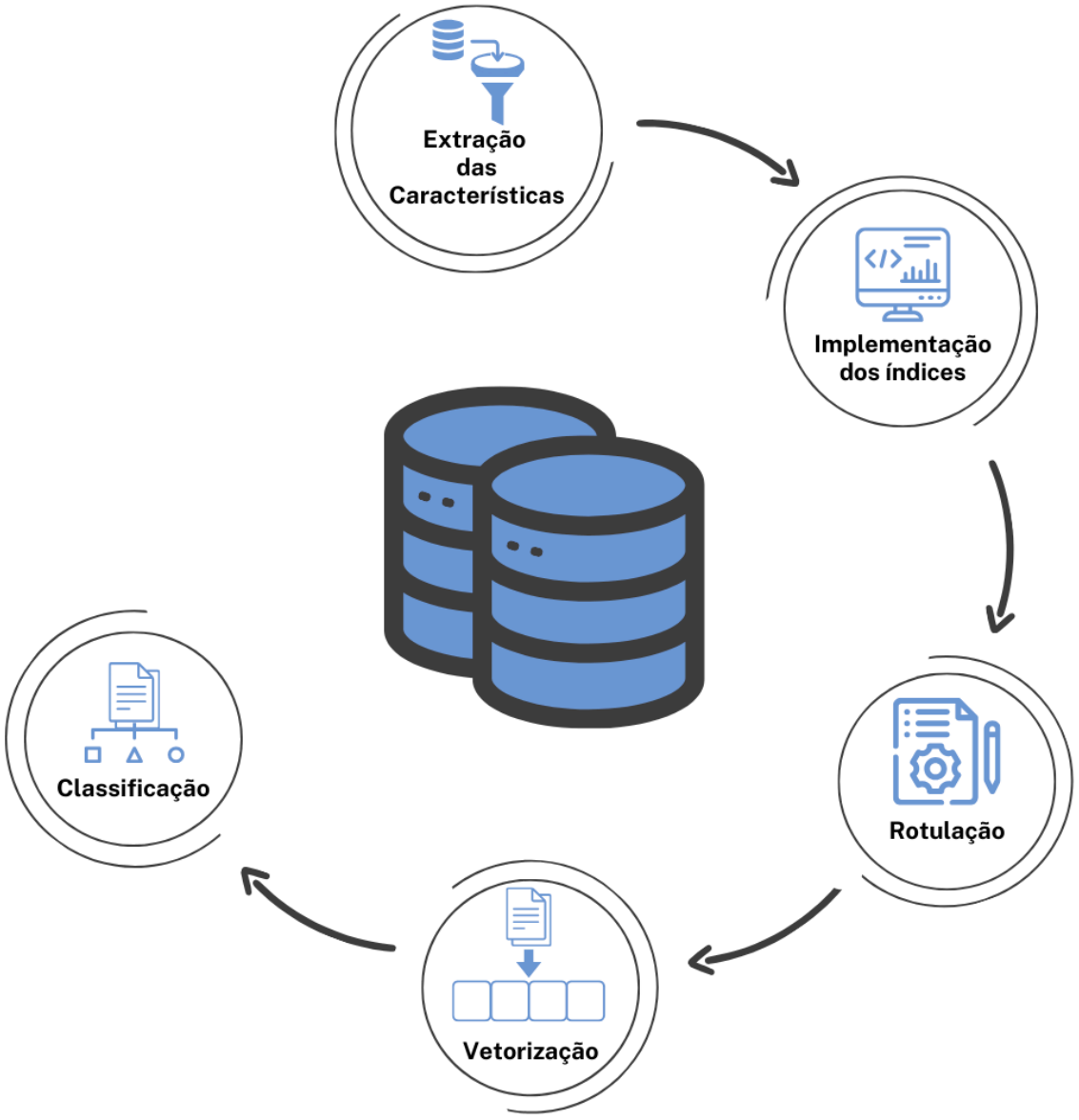
Schäfer, F., Zeiselmair, C., Becker, J., and Otten, H. (2018). Synthesizing crisp-dm and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions*.

Srisunakrua, T. and Chumworatayee, T. (2019). Readability of reading passages in english textbooks and the thai national education english test: A comparative study. *Arab World English Journal (AWEJ) Volume*, 10.

Tekfi, C. (1987). Readability formulas: An overview. *Journal of documentation*.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.

APÊNDICE A - FLUXOGRAMA DO PROCESSO DA MODELAGEM



APÊNDICE B - COMPLEMENTAÇÃO TEÓRICA E PRÁTICA

O conceito de leiturabilidade foi utilizado no estudo para avaliação qualitativa dos resultados, isto é, foram lidos os dados textuais com a nova rotulação de forma manual e verificou se estava condizente com os fatores da leiturabilidade (*e.g.,* Compreensão textual). Por outro lado, a legibilidade foi usada para avaliação qualitativa em que consistiu no uso das métricas, como índice de *Flesch-Kincaid* e índice de *Gunning Fog*. Além disso, foi utilizado na abordagem de extração de Características Estatísticas Textuais (CET). Com isso, possibilitou-se a avaliação quanti-quali de forma criteriosa com base na literatura científica.

Como o presente trabalho foi de caráter exploratório, pretende-se adaptar os métodos e modelos utilizados para aprimorar as etapas estabelecidas, assim, adquirindo resultados mais satisfatórios. Também, incluir especialistas de diversos domínios (*e.g.,* Profissionais da linguística) e utilizar a abordagem *Human-Centered Research (HCR)* com o intuito de enriquecer a pesquisa com conceitos e técnicas específicas do domínio para identificar nuances presentes no contexto e melhorar a experiência do usuário final com um *framework* que possa indicar trechos com uma linguagem mais complexa e utilizar de *Large Language Models* (*LLMs*), como Chat GPT, para simplificar ou adaptar o texto de acordo com o grau de entendimento do determinado utilizador do sistema.