Proposta recebida em Setembro 2024 e aceite para publicação em Novembro 2024.

Avaliação Automática do Nível de Complexidade de Textos em Português Europeu

Automatic Assessment of Text Complexity Levels in European Portuguese

Eugénio Ribeiro
INESC-ID Lisboa, Portugal
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

Nuno Mamede
INESC-ID Lisboa, Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal

Jorge Baptista
INESC-ID Lisboa, Portugal
Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal

Resumo

A avaliação da inteligibilidade de textos e a sua classificação por níveis de complexidade é essencial para o ensino de língua e para indústrias relacionadas com a linguagem que dependem de uma comunicação eficaz. O Quadro Europeu Comum de Referência para as Línguas (CEFR) é uma referência amplamente reconhecida para a classificação dos níveis de proficiência linguística. Este quadro pode ser utilizado não apenas para avaliar a proficiência de aprendentes de uma língua, mas também, de uma perspetiva de inteligibilidade, como um meio de identificar a proficiência necessária para compreender um texto. O objetivo deste estudo é desenvolver e avaliar modelos automáticos capazes de classificar textos em português europeu de acordo com os níveis de complexidade definidos pelo CEFR. Para tal, exploramos o ajuste de vários modelos de base pré-treinados em dados textuais utilizados para fins de avaliação de proficiência e exploramos abordagens que tiram partido da natureza ordinal dos níveis. Realizamos ainda uma análise preliminar da capacidade de base que modelos baseados em instruções têm para desempenhar esta tarefa. Nas experiências, os melhores modelos conseguem atingir mais de 80% de taxa de acerto e 75% de medida F_1 mas têm dificuldade em generalizar para diferentes tipos de texto, o que revela a necessidade de dados de treino adicionais e mais diversificados.

Palavras chave

inteligibilidade; complexidade textual; português europeu

Abstract

The assessment of text readability and the classification of texts by complexity levels is essential for language education and language-related industries that rely on effective communication. The Common European Framework of Reference for Languages (CEFR) provides a widely recognized framework for classifying language proficiency levels. This framework can be used not only to assess the proficiency of learners of a given language, but also from a readability perspective, as a means to identify the proficiency required to understand specific pieces of text. This study aims to develop and evaluate automatic models capable of classifying texts in European Portuguese according to the complexity levels defined by the CEFR. For that, we explore the fine-tuning of several foundation models on textual data used for proficiency evaluation purposes. Additionally, we explore approaches to leverage the information provided by the ordinal nature of the levels. Furthermore, we perform a preliminary analysis of the base capability of instruction-based models to perform this task. Our experiments show that the best models can achieve over 80% accuracy and 75% F_1 score. However, they have difficulty in generalizing to different types of text, which reveals the need for additional and more diverse training data.

Keywords

readability; text complexity; European Portuguese

DOI: 10.21814/lm.16.2.449

1. Introdução

Identificar o nível de inteligibilidade, complexidade, ou dificuldade de um texto é relevante em diversos domínios, abrangendo não só a aprendizagem de línguas, mas também várias indústrias relacionadas com a linguagem e muitas outras atividades humanas. Num contexto educacional, avaliar corretamente o nível de complexidade permite que educadores escolham textos adequados às capacidades dos alunos, promovendo um desenvolvimento linguístico eficaz e experiências de aprendizagem personalizadas. Além disso, fora do domínio da educação, a classificação do nível de inteligibilidade encontra aplicações em diferentes setores. Por exemplo, na indústria bancária, apresentar informações e políticas financeiras com um nível de complexidade apropriado garante que os clientes possam entender os termos e condições, levando à toma de decisões bem informadas. Da mesma forma, na área da saúde, é crucial para indivíduos com diferentes níveis de proficiência na língua que as instruções médicas, formulários de consentimento e materiais de informação sejam acessíveis e compreensíveis. Além disso, informações legais, comunicações governamentais e manuais de utilizador, entre muitos outros, podem beneficiar de uma avaliação adequada do nível de inteligibilidade, facilitando a comunicação, a transparência do conteúdo e a compreensão em geral.

No contexto da educação, o Quadro Europeu Comum de Referência para as Línguas (CEFR) (Council of Europe, 2001) define um enquadramento amplamente reconhecido para classificar níveis de proficiência linguística, variando entre A1 (iniciante) e C2 (proficiente). Além disso, tem duas adaptações específicas para o português europeu: o Quadro de Referência para o Ensino Português no Estrangeiro (QuaREPE) (Grosso et al., 2011) e o Referencial Camões PLE (Direção de Serviços de Língua e Cultura, 2017). Embora este quadro seja tipicamente usado para avaliar o nível de proficiência dos aprendentes de uma determinada língua, ele também pode ser visto de uma perspetiva de inteligibilidade, como um indicador do nível de proficiência necessário para entender um determinado texto (Hulstijn, 2007, 2011). Deste modo, explorando a perspetiva de inteligibilidade do CEFR, é possível analisar os fatores que influenciam a compreensão do texto e as suas implicações para a educação e para as indústrias relacionadas com a língua que procuram transmitir informações a aprendentes ou clientes de forma clara, concisa e de fácil compreensão.

Determinar o nível de complexidade ou inteligibilidade de textos apresenta seu próprio conjunto de desafios, especialmente ao trabalhar com línguas para as quais a quantidade de recursos anotados é limitada. Anotar grandes quantidades de dados de texto com os níveis definidos pelo CEFR é uma tarefa intensiva, demorada e que, na maioria das vezes, exige conhecimento especializado do domínio. Consequentemente, existe uma escassez de dados anotados, o que dificulta o desenvolvimento de modelos robustos e precisos para a classificação automática do nível de complexidade em múltiplos idiomas. Ainda assim, neste estudo, abordamos essa tarefa no contexto do português europeu.

Neste contexto, os dados usados na maioria dos estudos anteriores foram obtidos através da extração de textos usados nos exames para avaliação de proficiência pelo Camões, I.P.[1], o instituto oficial para a promoção da língua portuguesa. No entanto, esses dados não estão disponíveis publicamente, variam ao longo do tempo e, como demonstrado em estudos anteriores (Branco et al., 2014a; Curto, 2014), existem alguns problemas ao nível da classificação. Numa tentativa de mitigar alguns destes problemas, definimos um novo conjunto de teste com base nos exames modelo disponibilizados publicamente na página do instituto (Ribeiro et al., 2024a). Neste estudo analisamos este conjunto de teste em mais detalhe, inclusive através da reanotação do nível de complexidade dos textos por peritos.

Neste estudo, o nosso objetivo principal é investigar métodos para a classificação automática do nível de complexidade de textos em português europeu, fornecendo uma abordagem para a avaliação automática da inteligibilidade textual. Para alcançar este objetivo, neste artigo, agregamos e estendemos as análises dos nossos artigos publicados nas atas da conferência PROPOR 2024 (Ribeiro et al., 2024a,b), em que comparamos o desempenho de vários modelos de base existentes para o português quando ajustados para esta tarefa, assim como o desempenho de diferentes abordagens para capturar a relação ordinal entre os vários níveis. Além disso, tendo em conta a proliferação de modelos generativos baseados em instruções e o seu desempenho em várias tarefas de Processamento de Língua Natural (PLN) (Ouyang et al., 2022), fazemos também uma análise do desempenho destes modelos nesta tarefa e num contexto *zero-shot*.

Neste artigo, começamos por dar, na Secção 2, uma visão geral dos trabalhos relacionados sobre a avaliação automática do nível de comple-

[1] https://www.instituto-camoes.pt/

xidade de textos, com foco no português europeu. Em seguida, na Secção 3, analisamos o conjunto de dados extraído dos exames do Camões, I.P., com especial detalhe no novo conjunto de teste. Na Secção 4, descrevemos as abordagens que exploramos para prever automaticamente o nível de complexidade. A Secção 5 descreve a configuração experimental, incluindo as metodologias de avaliação. Em seguida, na Secção 6, apresentamos e discutimos os resultados das experiências, incluindo os erros e enviesamentos observados para os diferentes modelos. Por fim, na Secção 7, resumimos as contribuições deste estudo, discutimos as suas limitações e fornecemos diretrizes para investigação futura.

2. Trabalho Relacionado

A avaliação da inteligibilidade ou complexidade de textos é um problema que tem sido amplamente explorado ao longo dos anos (Leal & Aluísio, 2024). Tradicionalmente, o problema é abordado através da criação de fórmulas ou índices de inteligibilidade com base em informações estatísticas e/ou conhecimento de domínio (DuBay, 2004; Crossley et al., 2017). Entre estes, os mais usados são o Índice de Facilidade de Leitura de Flesch e o Índice de Legibilidade de Flesch-Kincaid (Kincaid et al., 1975).

No entanto, considerando os desenvolvimentos em aprendizagem automática, e especialmente em PLN, a investigação em avaliação automática da inteligibilidade e em tarefas relacionadas, como a avaliação da complexidade lexical (North et al., 2023), começou também a seguir as tendências na área de PLN. As primeiras abordagens (e muitas das iniciativas recentes para línguas com recursos limitados) basearam-se em características selecionadas manualmente, como a frequência de palavras, comprimento de frases e complexidade sintática, combinadas com algoritmos clássicos de aprendizagem automática, como árvores de decisão e Máquinas de Vetores de Suporte (SVM) (e.g., Aluisio et al., 2010; François & Fairon, 2012; Karpov et al., 2014; Curto et al., 2015; Pilán & Volodina, 2018; Forti et al., 2020; Leal et al., 2023). Posteriormente, surgiram abordagens de aprendizagem profunda baseadas em *embeddings* de palavras pré-treinados, como os gerados usando Word2Vec (Mikolov et al., 2013) (e.g., Cha et al., 2017; Nadeem & Ostendorf, 2018; Filighera et al., 2019). Por fim, mais recentemente, a investigação na área mudou-se para o ajuste de modelos pré-treinados baseados em transformadores, como BERT (Devlin et al., 2019), GPT (Radford et al., 2019) e RoBERTa (Liu et al., 2019); por exemplo, (Santos et al., 2021; Yancey et al., 2021; Martinc et al., 2021; Mohtaj et al., 2022).

Embora vários estudos tenham abordado a avaliação da inteligibilidade ou complexidade de textos como uma tarefa de regressão (e.g., Marujo et al., 2009; Cha et al., 2017; Nadeem & Ostendorf, 2018; Martinc et al., 2021; Wilkens et al., 2022; Mohtaj et al., 2022), apenas alguns exploraram as diferenças entre abordagens de regressão e classificação para a tarefa. Heilman et al. (2008) compararam regressão linear, regressão logística ordinal (McCullagh, 1980) e regressão logística multiclasse. A segunda abordagem obteve o melhor desempenho em termos de correlação, erro quadrático médio e taxa de acerto adjacente num cenário de validação cruzada. Contudo, o modelo de regressão linear mais simples revelou maior capacidade de generalização. Aluisio et al. (2010) compararam o desempenho de SVMs treinadas para classificação, regressão e classificação ordinal. Os modelos revelaram desempenho semelhante, mas cada um teve uma leve vantagem em termos de uma das métricas de avaliação, com a classificação a alcançar o maior valor de F_1, a regressão a maior correlação e a classificação ordinal o menor erro. Por fim, Xia et al. (2016) compararam uma abordagem de classificação usando SVMs com uma abordagem baseada em *ranking* de pares e alcançaram uma melhor correlação com a abordagem de classificação.

Tal como para a maioria das tarefas de PLN, uma parte significativa da investigação sobre avaliação automática do nível de inteligibilidade de texto foca-se na língua inglesa (e.g., Xia et al., 2016; Cha et al., 2017; Nadeem & Ostendorf, 2018; Filighera et al., 2019; Martinc et al., 2021). No entanto, neste caso, também existem vários estudos que abordam o problema em outras línguas, muitas das quais com recursos limitados. Por exemplo, existem estudos em francês (e.g., François & Fairon, 2012; François et al., 2020; Yancey et al., 2021; Wilkens et al., 2022; Hernandez et al., 2022), chinês (e.g., Sung et al., 2015), alemão (e.g., Mohtaj et al., 2022), italiano (e.g., Forti et al., 2020; Santucci et al., 2020), russo (e.g., Karpov et al., 2014; Reynolds, 2016), sueco (e.g., Jönsson et al., 2018; Pilán & Volodina, 2018) e esloveno (e.g., Martinc et al., 2021).

Focando no português, embora existam alguns estudos sobre a variedade brasileira da língua (e.g., Scarton & Aluísio, 2010; Aluisio et al., 2010; Leal et al., 2023), neste estudo estamos principalmente interessados na variedade europeia. Assim, apresentamos abaixo com mais detalhe estudos anteriores que cobrem esta variedade.

A versão portuguesa do sistema de tutoria REAP (Marujo et al., 2009) incluía um classificador de nível de inteligibilidade treinado em manuais escolares do 5º ao 12º ano. O modelo baseava-se em SVMs aplicados a características lexicais, como estatísticas de unigramas de palavras, e usava regressão logística ordinal para capturar a natureza ordinal dos níveis. Embora este modelo fosse preciso quando aplicado a manuais escolares, o seu desempenho diminuiu significativamente quando aplicado a exames do 6º, 9º e 12º ano de escolaridade.

A ferramenta LX-CEFR (Branco et al., 2014b) foi desenhada para ajudar aprendentes e professores de português a avaliar o nível de complexidade de um texto de acordo com o CEFR. A ferramenta foca-se em quatro características de forma independente: o Índice de Facilidade de Leitura de Flesch, a densidade lexical na forma da proporção de substantivos, o comprimento médio das palavras em número de sílabas e o comprimento médio das frases em número de palavras. Um corpus de 114 excertos extraídos dos exames de português como língua estrangeira realizados pelo Camões, I.P., foi usado para calcular a correlação entre estas características e o nível de inteligibilidade. Um estudo subsequente (Branco et al., 2014a) focou-se na reavaliação da ferramenta por especialistas humanos, assim como na reanotação dos textos por múltiplos instrutores de línguas. Neste caso, a concordância entre anotadores foi de apenas 0,17, o que revela a dificuldade e subjetividade da tarefa.

Curto et al. (2015) exploraram o uso de vários algoritmos clássicos de aprendizagem automática para a tarefa. Os algoritmos foram aplicados sobre 52 características divididas em 5 grupos diferentes: partes do discurso, blocos, frases e palavras, verbos, médias e frequências, e extras. As experiências foram realizadas numa versão estendida do conjunto de dados usado no contexto da ferramenta LX-CEFR, contendo 237 excertos. O melhor desempenho foi alcançado usando o algoritmo LogitBoost (Friedman et al., 2000). As características que se revelaram mais importantes foram as relacionadas com o tamanho dos textos em número de palavras e frases, o número de palavras diferentes, a quantidade de dependências sintáticas e o número de nós na árvore sintática. Além disso, à semelhança do que foi observado por Branco et al. (2014a), uma reanotação desta versão estendida do conjunto de dados por dois grupos de múltiplos especialistas revelou valores baixos para a concordância entre anotadores: 0,19 e 0,16 (Curto, 2014).

Akef et al. (2024) também exploraram o uso de algoritmos clássicos de aprendizagem, desta vez aplicados sobre um conjunto de 489 características separadas em 6 grupos: contagens, características lexicais, sintáticas, discursivas, morfológicas e psicolinguísticas. Além disso, os modelos foram treinados em múltiplas variantes do conjunto de dados extraído dos exames do Camões, I.P., incluindo uma com 500 excertos. O desempenho foi semelhante ao atingido por Curto et al. (2015). Contudo, experiências adicionais mostraram que os modelos têm pouca capacidade de generalização para um conjunto de dados composto por textos extraídos de livros para aprendizagem de português como segunda língua.

Correia & Mendes (2021) exploraram o uso de uma rede neuronal híbrida com dois ramos, um deles focado no processamento bidirecional de *embeddings* de palavras pré-treinados (NILC-Embeddings (Hartmann et al., 2017)) e o outro num conjunto de 14 características baseadas em comprimento e índices de inteligibilidade. A aplicação desta rede neuronal sobre o conjunto de dados do Camões, I.P., com 500 excertos mostrou que a combinação dos dois ramos leva a um desempenho superior àquele atingido por cada ramo individualmente.

Por fim, passando para o uso de abordagens baseadas no ajuste de modelos pré-treinados, Santos et al. (2021) exploraram o ajuste de versões em português dos modelos GPT-2 (Radford et al., 2019) e RoBERTa (Liu et al., 2019) em múltiplas variantes do conjunto de dados extraído dos exames do Camões, I.P.. Em geral, nas versões maiores do conjunto de dados, estas abordagens atingiram melhor desempenho que as abordagens clássicas, sendo os melhores resultados obtidos usando o modelo GPT-2. Akef et al. (2024) também exploraram o uso de uma versão ajustada do modelo GPT-3.5-Turbo, reportando um desempenho superior ao do modelo GPT-2, embora com uma metodologia de avaliação diferente. Nos nossos estudos (Ribeiro et al., 2024a,b), explorámos o uso de vários modelos de base adicionais, atingindo os melhores resultados usando um modelo da família Albertina PT-* (Rodrigues et al., 2023) e uma abordagem que tem em conta a relação ordinal entre os vários níveis. Como este artigo descreve uma extensão desses estudos, os resultados concretos serão apresentados e discutidos mais à frente. Relativamente aos resultados dos restantes estudos, tendo em conta as variações no conjunto de dados e na metodologia de avaliação, é difícil fazer uma comparação que cubra todas as abordagens. Ainda assim, a Tabela 1 sumariza os resultados

Abordagem	**Taxa de Acerto**	**Macro F_1**
LogitBoost (Curto et al., 2015)	68,60	64,30
LogitBoost (Akef et al., 2024)	68,00	58,66
Floresta Aleatória (Akef et al., 2024)	72,00	61,17
Rede Neuronal Híbrida (Correia & Mendes, 2021)	73,00	-
GPT-2 (Santos et al., 2021)	75,62	68,90
RoBERTa (Santos et al., 2021)	72,50	58,90
GPT-3.5-Turbo (Akef et al., 2024)	79,00	70,11

Tabela 1: Resultados obtidos em estudos anteriores sobre o conjunto de dados com 500 excertos extraídos dos exames do Camões, I.P.. Os resultados estão na forma de percentagem. Os resultados da abordagem de Curto et al. (2015) sobre este conjunto de dados foram obtidos por Santos et al. (2021). Os resultados do modelo GPT-3.5-Turbo foram obtidos sobre uma partição de teste enquanto os das restantes abordagens foram obtidos usando validação cruzada.

obtidos pelas abordagens aplicadas sobre o conjunto de dados com 500 excertos extraídos dos exames do Camões, I.P., em termos de taxa de acerto e medida macro F_1.

3. Conjunto de Dados

De modo semelhante aos estudos anteriores sobre a análise automática da complexidade de textos em português europeu discutidos na Secção 2, o nosso conjunto de dados é composto por textos extraídos dos exames para avaliação de proficiência realizados pelo Camões, I.P.[2], o instituto oficial de promoção da língua portuguesa. Os textos abrangem os níveis A1 a C1 do CEFR, conforme definidos na versão portuguesa do quadro (Grosso et al., 2011; Direção de Serviços de Língua e Cultura, 2017). Tendo em conta que estes textos são utilizados para fins de avaliação e podem ser reutilizados ao longo do tempo, eles não estão disponíveis publicamente, o que dificulta a investigação sobre esta tarefa. Por uma questão de simplicidade, doravante chamaremos estes exames de exames privados. Além disso, o número de textos anotados aumenta ao longo do tempo e não existe uma partição padrão dos dados, o que levou à utilização de múltiplas versões diferentes do conjunto de dados nos estudos anteriores. Como tal, é difícil comparar as abordagens existentes sem repetir as experiências. Por fim, alguns estudos (Branco et al., 2014a; Curto, 2014) demonstraram que existe discrepância entre os níveis dos textos e os que lhe são atribuídos por anotadores humanos. Em parte, esta discrepância pode ser explicada pela inferência do nível dos textos a partir do nível do exame do qual são extraídos. Uma vez que, no contexto dos exames, cada texto está associado a uma determinada tarefa a realizar, este nível é apenas uma aproximação. Ainda assim, a baixa concordância entre anotadores observada em estudos anteriores revela que parte da discrepância também se deve à subjetividade inerente à definição do nível de complexidade dos textos.

Numa tentativa de mitigar alguns destes problemas, definimos um novo conjunto de teste[3] com base nos textos incluídos nos exames modelo (um para cada nível) disponibilizados na página do instituto. Uma vez que estes exames estão disponíveis publicamente, este conjunto de teste pode ser usado para padronizar a avaliação, permitindo que investigadores sem acesso aos exames privados possam pelo menos avaliar as suas abordagens. Inicialmente, tal como para os textos extraídos dos exames privados, usamos o nível do exame como uma aproximação do nível dos textos. No entanto, tendo em conta a discrepância entre esta aproximação e o nível atribuído por anotadores humanos observada em estudos anteriores, optámos por também recorrer a peritos para fazer a anotação dos textos. A concordância (κ de Fleiss) entre os três anotadores foi de apenas 0,21, o que confirma a subjetividade e a dificuldade da tarefa. A concordância (κ de Cohen) entre o nível do exame e o nível médio atribuído pelos anotadores é de 0,29.

A Tabela 2 mostra a distribuição dos textos pelos níveis do CEFR. Aquando da realização deste estudo, havia 598 textos extraídos dos exames que não estão disponíveis publicamente. Podemos ver que há um viés na direção do nível central (B1) e um número menor de exemplos dos níveis mais avançados. Além disso, considerando que alguns dos textos são reutilizados em

[2] https://www.instituto-camoes.pt/

[3] https://gitlab.hlt.inesc-id.pt/iread4skills/cmodel

	A1	A2	B1	B2	C1	Total
Treino	92	157	240	49	60	598
Teste (E)	8	12	5	3	4	32
Teste (A)	6	14	4	7	1	32

Tabela 2: Distribuição dos textos extraídos dos exames do Camões, I.P., pelos níveis do CEFR. As anotações (E) e (A) associadas ao conjunto de teste correspondem ao nível do exame e ao nível médio atribuído pelos anotadores, respetivamente.

vários exames, alguns dos exemplos consistem em pequenas variações do mesmo texto.

Relativamente ao conjunto de teste extraído dos exames modelo, podemos ver que este consiste em apenas 32 textos, o que introduz algumas limitações. Os textos cobrem vários tipos, como excertos de narrativas, diálogos, notícias e cartas. O Apêndice A inclui um exemplo de cada nível, selecionado de entre os casos em que existe concordância entre o nível do exame e o atribuído pelos anotadores. Neste contexto, analisando as distribuições na Tabela 2, podemos ver que existe uma tendência dos anotadores para evitar o nível C1. No entanto, em geral, a variação entre as duas classificações foi equilibrada (7 textos considerados de um nível mais fácil pelos anotadores e 8 de um nível mais avançado) e em apenas um caso foi superior a um nível. Em comparação com a distribuição dos textos dos exames privados, podemos ver que no conjunto de teste existe uma prevalência maior dos níveis A. Isto deve-se a um tipo de exercício de compreensão escrita que envolve múltiplos textos curtos e que só ocorre nos exames desses níveis. A Tabela 3 mostra alguns exemplos deste tipo de textos. A diferença em relação à distribuição dos textos extraídos dos exames privados deve-se à não inclusão deste tipo de textos no conjunto disponibilizado, que consiste apenas em textos mais longos. Como tal, o conjunto de teste extraído dos exames modelo serve também para avaliar a capacidade de generalização dos modelos para tipos de texto não observados durante o treino.

4. Abordagens

Analisando os resultados dos estudos anteriores sobre a avaliação automática do nível de complexidade de textos referidos na Secção 2, chegamos à conclusão de que, atualmente, o melhor desempenho na tarefa é atingido através do ajuste de modelos de base pré-treinados (Bommasani et al., 2021). Como tal, o principal objetivo deste estudo é comparar o desempenho dos diversos modelos de base existentes para o português quando ajustados para esta tarefa. No entanto, tendo em conta a proliferação de modelos generativos baseados em instruções e o seu desempenho em várias tarefas de PLN (Ouyang et al., 2022), fazemos também uma análise preliminar do desempenho destes modelos para esta tarefa.

4.1. Ajuste de Modelos Pré-Treinados

O ajuste de modelos pré-treinados, conhecido como *fine-tuning*, é uma técnica na qual um modelo previamente treinado numa grande quantidade de dados é adaptado para uma tarefa específica. Isso é feito continuando o treino do modelo, mas com um conjunto de dados mais especializado e um objetivo de tarefa particular, como a classificação de textos por nível de complexidade.

Apesar de os níveis do CEFR terem uma natureza ordinal, os estudos mais recentes sobre a avaliação automática do nível de complexidade de textos em português (Curto et al., 2015; Santos et al., 2021; Correia & Mendes, 2021; Akef et al., 2024) abordaram o problema como uma tarefa de classificação sem considerar as relações entre os níveis. Posto isto, neste estudo, além de analisar o desempenho dos vários modelos de base quando ajustados para a tarefa de classificação, exploramos também abordagens que tentam tirar partido da natureza ordinal dos níveis.

Em primeiro lugar, dado o mesmo modelo de base, comparamos o seu desempenho quando ajustado para a tarefa de classificação com o seu desempenho quando ajustado para uma tarefa de regressão. No primeiro caso, cada nível é considerado uma classe independente enquanto, no segundo, os níveis são convertidos em valores numéricos ordenados. No caso da classificação, o modelo produz uma aproximação da distribuição de probabilidade das classes usando a função *softmax*, sendo escolhida aquela com maior probabilidade. No caso da regressão, o modelo produz um valor numérico contínuo. Neste caso, a previsão do nível é obtida por arredondamento ao nível mais próximo.

Além disso, exploramos a adaptação *a posteriori* do modelo de classificação para ter em conta a ordem dos níveis usando duas abordagens distintas. A primeira consiste em calcular a média ponderada da aproximação da distribuição de probabilidades dada pela função *softmax*. A segunda consiste em usar essa distribuição de probabilidades para calcular a função de distribuição acumulada que, neste caso, representa a probabilidade de a complexidade de um determinado texto ser menor ou igual a um determinado nível. Em seguida, usamos essa função para treinar um modelo de regressão logística ordinal.

Exame	Anotadores	Exemplo
A1	A1	*É proibido sair da sala sem autorização do professor.*
A1	A2	*É favor não jogar à bola no interior da escola.*
A2	A1	*Avariado. Pedimos desculpa pelo incómodo.*
A2	A2	*Lamentamos mas não é possível atendê-lo agora. Tente mais tarde.*

Tabela 3: Exemplos de textos curtos que ocorrem apenas nos exames dos níveis A.

4.2. Modelos Baseados em Instruções

Recentemente, temos observado a proliferação de modelos generativos baseados em instruções (Ouyang et al., 2022), tendo o ChatGPT (OpenAI, 2023) como principal impulsionador. Estes modelos têm revelado um desempenho satisfatório em várias tarefas de PLN, especialmente em contextos com poucos ou mesmo nenhuns exemplos. Logo, é importante avaliar o desempenho destes modelos na análise do nível de complexidade de textos.

Neste estudo, focamo-nos apenas num cenário *zero-shot*, ou seja, não são fornecidos quaisquer exemplos ao sistema. Deste modo, analisamos a capacidade de base destes modelos e o seu conhecimento intrínseco para desempenhar esta tarefa. Os cenários que envolvem exemplos introduzem níveis de complexidade adicionais, relacionados com a escolha dos exemplos a fornecer e com as limitações ao nível do tamanho da janela de contexto que os modelos são capazes de processar. Como tal, deixamos a exploração desses cenários para trabalho futuro.

O desempenho dos modelos baseados em instruções é significativamente influenciado pela qualidade das instruções (Mishra et al., 2022; Giray, 2023; Chen et al., 2023). Neste contexto, exploramos o uso de duas instruções diferentes:

Instrução A Classifica o nível de complexidade do seguinte texto de acordo com o Quadro Europeu Comum de Referência para as Línguas (CEFR). Responde apenas com o nível: A1, A2, B1, B2, C1.

Instrução B De acordo com o Quadro Europeu Comum de Referência para as Línguas (CEFR), qual é o nível de proficiência necessário para compreender o seguinte texto? Responde apenas com o nível: A1, A2, B1, B2, C1.

Enquanto a primeira instrução aponta diretamente ao nível de complexidade do texto, a segunda aborda o problema do ponto de vista do nível de proficiência necessário para compreender o texto.

5. Configuração Experimental

Esta secção descreve a nossa configuração experimental, começando com a abordagem usada como linha de base para as nossas experiências na Secção 5.1. Em seguida, descrevemos os modelos usados nas nossas abordagens na Secção 5.2. A Secção 5.3 descreve a metodologia de avaliação. Por fim, a Secção 5.4 fornece mais alguns detalhes de implementação para aumentar a reprodutibilidade das nossas experiências.

5.1. Linha de Base

Como linha de base usamos um classificador baseado em *gradient boosting* (Friedman, 2001) aplicado a um conjunto de 631 características descritivas, lexicais, sintáticas e discursivas, que foram identificadas no contexto do projeto iRead4Skills[4] como potenciais indicadores da complexidade de um texto[5].

5.2. Modelos

As abordagens usadas no nosso estudo tiram partido de modelos existentes, seja para os ajustar à tarefa ou para os usar diretamente no caso da abordagem baseada em instruções. Nesta secção, descrevemos os vários modelos usados neste estudo, começando pelos modelos de base pré-treinados, na Secção 5.2.1, seguidos dos modelos baseados em instruções, na Secção 5.2.2.

5.2.1. Modelos de Base

Em termos de modelos de base pré-treinados (Bommasani et al., 2021), o nosso objetivo é cobrir extensivamente os modelos que estão atualmente disponíveis publicamente para o português, independentemente da variedade linguística (brasileira ou europeia). Estes modelos são descritos em seguida:

BERTimbau (Souza et al., 2020) é o modelo de base em português mais utilizado. Este segue

[4] https://iread4skills.com/

[5] https://gitlab.hlt.inesc-id.pt/iread4skills/docs

a arquitetura original do BERT (Devlin et al., 2019), mas foi treinado no Brazilian Web as a Corpus (brWaC) (Wagner Filho et al., 2018) exclusivamente para modelação de língua com *tokens* mascarados. Inicialmente foram treinadas duas variantes do modelo: *base* e *large* com 110M e 335M de parâmetros, respetivamente. Existe também uma versão destilada, obtida através da aplicação da abordagem DistilBERT (Sanh et al., 2019) sobre a variante *base*.

BERTugues (Zago, 2023) é um modelo que foi treinado com o objetivo de superar o BERTimbau. Este foi treinado numa versão filtrada do brWaC em que textos com pouca qualidade foram descartados. Além disso, também foi treinado para previsão da próxima frase. O seu tokenizador inclui *emojis* e descarta caracteres que ocorrem apenas muito raramente em português. Contudo, ao contrário do BERTimbau, o BERTugues só possui uma variante *base*, com 110M de parâmetros.

RoBERTa PT (Santos et al., 2021) é uma versão pequena do RoBERTa (Liu et al., 2019) com 68M de parâmetros, treinada em 10 milhões de frases em português e 10 milhões de frases em inglês do corpus OSCAR (Suárez et al., 2019). Foi treinado por Santos et al. (2021) para ser usado num estudo sobre avaliação automática do nível de inteligibilidade.

GPorTuguese-2 (Guillou, 2020) é uma versão ajustada do modelo GPT-2 *small* em inglês (Radford et al., 2019) na Wikipedia em português. Possui 124M de parâmetros. Este foi o modelo usado como base para alcançar o melhor desempenho no estudo sobre avaliação automática do nível de inteligibilidade de Santos et al. (2021).

Albertina PT-* (Rodrigues et al., 2023; Santos et al., 2024) é uma família de modelos baseados no DeBERTa (He et al., 2021). Existem modelos para o português europeu e para o português do Brasil. Para cada variedade linguística, existem três variantes, com 100M, 900M e 1.5B de parâmetros, respetivamente. Os modelos para português do Brasil foram treinados no brWaC, enquanto os modelos para português europeu foram treinados numa combinação de transcrições de debates no Parlamento Português, as partes em português dos corpora do Parlamento Europeu e a parte em português europeu do corpus OSCAR. Atualmente, estes modelos pré-treinados são os que atingem o melhor desempenho quando ajustados para múltiplas tarefas de PLN em português.

5.2.2. Modelos Baseados em Instruções

Relativamente aos modelos baseados em instruções, tal como descrito na Secção 4.2, o nosso objetivo é fazer um estudo preliminar da capacidade destes modelos para desempenhar a tarefa de classificação da complexidade de textos num cenário *zero-shot*. Tendo em conta que o ChatGPT (OpenAI, 2023) é o principal impulsionador do uso de modelos baseados em instruções, exploramos o uso de vários modelos da família GPT (Radford et al., 2018). No entanto, tendo em conta que esses modelos não são abertos, exploramos também o uso de modelos das famílias Llama (Touvron et al., 2023) e Mistral (Jiang et al., 2023).

GPT (Radford et al., 2018) é uma família de modelos que conta já com múltiplas gerações, todas elas com impacto significativo na área de PLN. A segunda geração desta família (Radford et al., 2019) já foi referida anteriormente como a base do modelo GPorTuguese-2 (Guillou, 2020). Porém, neste contexto, estamos interessados nos modelos baseados em instruções que surgiram após o desenvolvimento do InstructGPT (Ouyang et al., 2022) e que estão na base do ChatGPT (OpenAI, 2023). Mais especificamente, nas nossas experiências exploramos o uso dos seguintes modelos: GPT 3.5 Turbo, GPT 4 Turbo (OpenAI, 2024), GPT 4o e GPT 4o Mini. Embora se saiba que a capacidade geral destes modelos evoluiu entre gerações e que alguns dos modelos têm um tamanho mais reduzido (GPT 3.5 Turbo e GPT 4o Mini), os modelos são privados e, como tal, os detalhes da sua implementação não são conhecidos.

Llama (Touvron et al., 2023) é uma família de modelos disponíveis publicamente que foi desenvolvida para competir com os modelos da família GPT. A geração atual, Llama 3 (Llama Team, 2024), inclui modelos com 8B, 70B e 405B de parâmetros. Por uma questão de recursos, nas nossas experiências exploramos apenas o uso do modelo com 8B de parâmetros.

Mistral (Jiang et al., 2023) é uma família de modelos originalmente desenvolvidos para serem mais eficientes do que os da família Llama, atingindo um desempenho semelhante com um número menor de parâmetros. No entanto, a razão para explorarmos modelos desta família é o subconjunto de modelos Mixtral (Jiang et al., 2024), baseados em Mistura Esparsa de Especialistas (SMoE). Esta abordagem combina vários conjuntos de parâmetros (especialistas), sendo apenas um subconjunto deles selecionado em cada passo de geração. Existem dois modelos

deste tipo: Mixtral 8x7B e Mixtral 8x22B. Ambos são compostos por oito especialistas, sendo a diferença o número de parâmetros de cada especialista. Tal como para a família Llama, por uma questão de recursos, exploramos apenas o uso do modelo mais pequeno, com 7B de parâmetros por especialista.

5.3. Metodologia de Avaliação

Começando pelas métricas de avaliação, damos principal relevância à taxa de acerto (TA), à taxa de acerto adjacente (TAA) e à medida macro F_1, que são algumas das métricas mais comuns em estudos anteriores sobre classificação automática do nível de complexidade. A taxa de acerto avalia a identificação exata do nível de complexidade de um texto, enquanto a taxa de acerto adjacente também considera níveis vizinhos, fornecendo informação sobre a dispersão das classificações pelos níveis adjacentes. Considerando que a distribuição dos textos pelos níveis não é equilibrada, a medida macro F_1 também é uma métrica relevante para perceber se os classificadores tendem a favorecer as classes maioritárias. Adicionalmente, na parte do nosso estudo em que a tarefa é abordada como um problema de regressão, também analisamos a Raíz do Erro Quadrático Médio (REQM). Além disso, pontualmente, utilizamos valores de precisão e sensibilidade para analisar o desempenho para níveis específicos.

Os estudos sobre avaliação automática do nível de complexidade em Português Europeu descritos na Secção 2 utilizaram abordagens de validação cruzada para a avaliação. Como mencionado por Santos et al. (2021), a validação cruzada não é uma prática comum ao treinar modelos neuronais de grandes dimensões, pois é um processo demorado. Contudo, embora tenhamos definido um novo conjunto de teste para avaliação, também recorremos a uma abordagem de validação cruzada em 10 partes para ajustar os hiperparâmetros e identificar os modelos de base pré-treinados que levam ao melhor desempenho quando ajustados para a tarefa. Isto permite-nos avaliar o desempenho dos nossos modelos num cenário de avaliação semelhante ao dos estudos anteriores e, em seguida, usar o conjunto de teste para avaliar a capacidade de generalização dos modelos com melhor desempenho.

Em cada parte do processo de validação cruzada, os modelos de base são ajustados durante 20 épocas. Os pesos da melhor época são então selecionados de acordo com a taxa de acerto do modelo. Considerando que o processo de validação cruzada gera 10 modelos ajustados diferentes para cada modelo de base, usamos uma combinação destes modelos para gerar as previsões para o conjunto de teste. Para agregar as previsões dos vários modelos, experimentámos abordagens baseadas em probabilidade, ranking e votação por maioria. Não conseguimos identificar uma abordagem que fosse claramente melhor do que as outras. Por isso, optámos por fazer a média das distribuições de probabilidade das classes previstas pelos vários modelos.

Para aumentar a robustez e mitigar o impacto da aleatoriedade, realizámos múltiplas execuções experimentais independentes para todos os modelos. Algumas das experiências realizadas nos nossos estudos anteriores (Ribeiro et al., 2024a,b) consideraram 10 execuções. No entanto, por uma questão de consistência e devido à quantidade de recursos necessária para realizar 10 execuções de todas as experiências, optámos por considerar apenas 3 execuções neste estudo. Como tal, alguns dos resultados poderão ser ligeiramente diferentes. No caso do ajuste de modelos pré-treinados, cada execução usa uma semente aleatória diferente para o processo de partição da validação cruzada. No caso dos modelos baseados em instruções, uma vez que não é feito nenhum processo de treino adicional, estes são aplicados apenas ao conjunto de teste. Além disso, ajustamos os parâmetros disponíveis de forma a minimizar a variabilidade das respostas dos modelos entre diferentes execuções.

Salvo se indicado explicitamente, as métricas de avaliação são reportadas como a média e o desvio padrão das várias execuções. Todas as métricas exceto a REQM são reportadas em forma de percentagem.

5.4. Detalhes de Implementação

Para treinar o classificador baseado em características usado como linha de base, utilizamos a implementação do algoritmo *gradient boosting* fornecida pela biblioteca scikit-learn (Pedregosa et al., 2011).

Para ajustar os modelos pré-treinados, utilizamos a funcionalidade oferecida pela biblioteca Transformers da HuggingFace (Wolf et al., 2020). Usamos os valores padrão para a maioria dos hiperparâmetros, com a exceção do tamanho do lote (*batch size*) e a taxa de aprendizagem, para os quais experimentamos vários valores. Para a maioria dos modelos de base, os melhores resultados foram obtidos usando lotes de tamanho 32 e uma taxa de aprendizagem de 5×10^{-5}. Uma das exceções é o GPorTuguese-2, cujo desempenho é altamente influenciado pelo preenchimento (*padding*). Portanto, para maximizar esse desempenho, utilizamos lotes com apenas

um exemplo. Para além disso, os melhores resultados foram obtidos com uma taxa de aprendizagem de 1×10^{-5}. A outra exceção refere-se às versões maiores dos modelos Albertina PT-*, com 900M e 1.5B de parâmetros, que exibiram comportamento errático para valores maiores de ambos os hiperparâmetros. Por isso, utilizamos lotes de tamanho 16 e uma taxa de aprendizagem de 1×10^{-5}.

Para aplicar a regressão logística ordinal, usamos a implementação do fornecida pela biblioteca statsmodels (Seabold & Perktold, 2010).

Por fim, relativamente aos modelos baseados em instruções, usamos a API da OpenAI[6] para aceder aos modelos da família GPT. Para os modelos disponíveis publicamente, utilizamos a funcionalidade oferecida pela biblioteca Transformers da HuggingFace (Wolf et al., 2020) e as versões dos modelos treinadas especificamente para seguir instruções. Caso a resposta dos modelos inclua várias classificações, consideramos apenas a primeira.

6. Resultados

Na Secção 6.1 começamos por apresentar e analisar o desempenho dos vários modelos ajustados no cenário de validação cruzada. Em seguida, na Secção 6.2, avaliamos a capacidade de generalização dos melhores modelos, analisando o seu desempenho no conjunto de teste. Por fim, na Secção 6.3, usamos esse mesmo conjunto de teste para analisar o desempenho dos modelos baseados em instruções.

6.1. Validação Cruzada

A primeira linha da Tabela 4 mostra os resultados que usamos como linha de base no nosso estudo, obtidos através da aplicação do algoritmo *gradient boosting* a um conjunto de 631 características descritivas, lexicais, sintáticas e de discurso. Neste caso, atingimos uma taxa de acerto de 75,31%, superior à atingida por Curto et al. (2015) e Akef et al. (2024) num conjunto de dados mais reduzido. Isto revela a importância de ter uma maior quantidade de dados anotados para treinar os modelos.

As restantes linhas da Tabela 4 mostram os resultados obtidos ao ajustar os múltiplos modelos de base para esta tarefa. Primeiro, podemos ver que todos os modelos atingiram uma taxa de acerto acima de 75%, superando a linha de base. Tal como observado no estudo de Santos et al. (2021), isto mostra que, quando existe um número suficiente de exemplos de treino, o ajuste de modelos pré-treinados leva a um melhor desempenho que os modelos obtidos usando técnicas clássicas de aprendizagem automática. Além disso, o melhor modelo no estudo de Santos et al. (2021) obteve uma taxa de acerto de 75,62% na versão do conjunto de dados com 500 excertos. Mais uma vez, isto mostra que os dados de treino adicionais disponíveis para o nosso estudo têm um impacto significativo no desempenho dos modelos.

Analisando modelos específicos, começando pelo BERTimbau, o modelo pré-treinado mais utilizado para o português, podemos ver que o desempenho das suas três variantes é o esperado, com o modelo *large* superando o modelo *base*, e a versão destilada trocando menos de 1% de desempenho por um tamanho reduzido e tempos de treino e inferência mais rápidos.

O BERTugues conseguiu superar a versão *large* do BERTimbau, apesar de ter o mesmo número de parâmetros da versão *base*. Isto também foi observado por seu autor em outras tarefas de PLN em português (Zago, 2023) e revela as vantagens de treinar modelos de língua em dados com qualidade superior e de ter um tokenizador mais adequado à língua focada.

O RoBERTa PT, o modelo mais pequeno usado no nosso estudo, atingiu um desempenho semelhante ao da versão *large* do BERTimbau em termos de taxa de acerto e medida F_1 e o melhor desempenho global em termos de taxa de acerto adjacente. Isto pode ser justificado pelas melhorias no processo de treino utilizado pelo RoBERTa, como, por exemplo, o uso de máscaras dinâmicas (Liu et al., 2019). Além disso, é expectável que o pré-treino em frases em português do corpus OSCAR também contribua para o desempenho, pois faz com que a variedade europeia da língua seja considerada.

O GPorTuguese-2, o único modelo da família GPT (Radford et al., 2019) utilizado nesta parte do nosso estudo, é um dos que apresenta melhor desempenho, ocupando o segundo lugar em ¡todas as métricas. Tal como observado por Santos et al. (2021), este modelo superou o RoBERTa PT em termos de taxa de acerto (por dois pontos percentuais em comparação com três no estudo de Santos et al. (2021)). O desempenho obtido com este modelo sugere que ele ainda é uma escolha segura, apesar da existência de modelos de base pré-treinados mais recentes. No entanto, como o uso de preenchimento (*padding*) tem um grande impacto no seu desempenho, não é possível aproveitar ao máximo o hardware mo-

[6] https://openai.com/api/

Modelo	**TA**	**TAA**	**F_1**
Linha de Base	75,31±0,57	94,42±0,52	66,74±0,90
BERTimbau Large	79,26±2,09	95,99±0,61	71,68±2,61
BERTimbau Base	78,26±1,67	95,71±0,59	71,30±2,60
BERTimbau Distilled	77,65±0,68	95,71±0,51	70,98±0,60
BERTugues	79,43±0,29	95,54±0,51	72,76±0,77
RoBERTa PT	79,15±0,75	**97,05±0,25**	71,49±1,15
GPorTuguese-2	81,16±0,63	96,71±0,92	74,81±1,60
Albertina PT-PT 1.5B	75,69±0,86	93,26±0,10	68,51±1,16
Albertina PT-BR 1.5B	75,64±1,06	93,48±0,73	68,85±2,84
Albertina PT-PT 900M	77,42±0,34	94,48±0,67	70,92±0,65
Albertina PT-BR 900M	76,15±0,59	93,42±0,82	69,07±0,70
Albertina PT-PT 100M	**81,77±0,44**	96,27±0,54	**76,17±1,01**
Albertina PT-BR 100M	80,43±1,60	95,99±0,61	73,88±1,67

Tabela 4: Resultados obtidos pelos modelos ajustados no cenário de validação cruzada.

derno durante o ajuste deste modelo, tornando-o mais lento do que o ajuste da variante *large* do BERTimbau e quase tão lento quanto o ajuste dos modelos da família Albertina PT-* com um número de parâmetros nove vezes superior.

Analisando os resultados dos modelos da família Albertina PT-*, podemos ver que os modelos pré-treinados em dados em português europeu superam seus equivalentes em português do Brasil, embora as diferenças sejam menos significativas no caso dos modelos maiores. Isto confirma que as diferenças entre as duas variedades são relevantes e têm impacto na forma como o nível de dificuldade de um texto é percebido.

Além disso, entre os modelos dessa família, podemos encontrar os modelos quer com melhor quer com pior desempenho nesta tarefa. Os modelos maiores, com 900M e 1.5B de parâmetros, que alcançam desempenho de ponta em várias tarefas de PLN em português, foram os que obtiveram os piores resultados no nosso estudo em todas as métricas. No entanto, podemos argumentar que se trata de um caso de sobreajuste, pois esses modelos são demasiado grandes para o número de exemplos de treino disponíveis. Assim, é expectável que eles apresentem melhor desempenho com uma quantidade suficientemente grande e representativa de dados de treino. Por outro lado, os modelos com 100M de parâmetros estão entre os que apresentam melhor desempenho na tarefa, alcançando uma taxa de acerto acima de 80%.

Globalmente, o melhor desempenho no cenário de validação cruzada foi obtido ao ajustar a versão do modelo Albertina PT-PT com 100M de parâmetros. A taxa de acerto foi de 81,77% e a medida F_1 foi de 76,17%. Isto também representa a menor diferença entre as duas métricas entre todos os modelos. A este respeito, Santos et al. (2021) observaram uma diferença de 13,60 pontos percentuais ao usar o RoBERTa PT e de 6,72 pontos percentuais ao usar o GPorTuguese-2. Esses valores foram reduzidos para 7,66 e 6,35 no nosso estudo, o que sugere que os dados de treino adicionais levam a modelos menos tendenciosos. Ainda assim, a diferença entre as métricas sugere que os modelos ainda são um pouco tendenciosos ou que, pelo menos, têm mais dificuldade em identificar exemplos de certos níveis.

A Tabela 5 mostra as matrizes de confusão das melhores execuções dos dois modelos com melhor desempenho em comparação com linha de base. Podemos ver que em todos os casos, a sensibilidade é de aproximadamente 90% para o nível B1, que é simultaneamente o nível intermédio e o mais proeminente no conjunto de dados de treino. Por outro lado, os modelos parecem ter algumas dificuldades em distinguir entre os níveis A. A principal diferença entre os dois melhores modelos parece ser a forma como eles lidam com os níveis avançados. Enquanto o GPorTuguese-2 parece ter algumas dificuldades em distinguir entre os níveis B2 e C1, o Albertina PT-PT, com 100M de parâmetros, parece ser mais tendencioso em direção à previsão do nível B1.

Passando para a comparação entre as abordagens de classificação e de regressão, as duas primeiras linhas da Tabela 6 mostram os resultados obtidos por versões do modelo Albertina PT-PT 100M ajustadas para cada uma das visões sobre o problema. As duas linhas restantes mostram os resultados das adaptações do modelo de classificação para tentar capturar a relação ordinal entre os níveis: a média ponderada da aproximação da distribuição de probabilidade dada

		Linha de Base					Albertina PT-PT 100M						GPorTuguese-2				
		Previsto					Previsto						Previsto				
		A1	**A2**	**B1**	**B2**	**C1**		**A1**	**A2**	**B1**	**B2**	**C1**		**A1**	**A2**	**B1**	**B2**	**C1**
Real	**A1**	62	26	4	0	0	**Real** **A1**	73	16	3	0	0	**Real** **A1**	73	17	2	0	0
	A2	18	125	14	0	0	**A2**	22	133	2	0	0	**A2**	29	127	1	0	0
	B1	1	13	215	5	6	**B1**	3	10	216	6	5	**B1**	3	7	218	6	6
	B2	0	2	14	18	15	**B2**	0	0	14	28	7	**B2**	0	0	6	28	15
	C1	0	1	15	9	35	**C1**	0	2	13	4	41	**C1**	0	0	3	15	42

Tabela 5: Matrizes de confusão das melhores execuções dos modelos com maior taxa de acerto no cenário de validação cruzada em comparação com a linha de base (76,09%): Albertina PT-PT 100M (82,11%) e GPorTuguese-2 (81,60%).

Abordagem	**REQM**	**TA**	**TAA**	**F_1**
Classificação	0,5341±0,0327	81,77±0,44	96,27±0,54	76,17±1,01
Regressão	**0,4940±0,0268**	80,77±0,58	**97,83±0,60**	75,33±1,12
Média *Softmax*	0,5134±0,0132	**82,33±0,26**	96,93±0,54	**76,74±0,60**
Regressão Ordinal	0,5439±0,0146	80,60±1,31	97,16±0,73	74,69±0,99

Tabela 6: Resultados obtidos pelas abordagens de classificação e regressão no cenário de validação cruzada.

pela função *softmax* e a aplicação de uma abordagem de regressão logística ordinal sobre essa distribuição. De forma semelhante ao observado por Aluisio et al. (2010), podemos ver que as diferenças de desempenho entre as abordagens são pequenas, nunca chegando aos dois pontos percentuais, exceto no caso da REQM, já que o modelo de regressão é treinado especificamente para minimizar essa métrica. Além disso, no nosso estudo original com dez execuções de treino (Ribeiro et al., 2024b), as diferenças foram ainda menores. Isto sugere que o modelo pré-treinado usado como base e os dados utilizados para o ajuste são mais relevantes do que capturar a natureza ordinal dos níveis de inteligibilidade. Ainda assim, é possível que as diferenças se possam tornar mais evidentes se forem considerados dados mais diversificados.

Comparando os resultados em termos de métricas específicas, como referido anteriormente, a abordagem de regressão apresenta uma REQM menor do que a abordagem de classificação. No entanto, isto ocorre às custas de um impacto de cerca de um ponto percentual ao nível da taxa de acerto e da medida F_1, o que sugere que o maior número de neurónios na camada de saída melhora a capacidade do modelo para capturar as características específicas de cada nível, o que pode ser utilizado para obter uma aproximação melhor do nível real de um texto. Por outro lado, o modelo de regressão tem o melhor desempenho quanto à taxa de acerto adjacente. Tal sugere que os erros do modelo de classificação são mais dispersos.

Analisando os resultados das adaptações do modelo de classificação para tentar capturar a relação ordinal entre os níveis, podemos ver que a aplicação da abordagem de regressão ordinal leva a uma melhoria do desempenho ao nível da taxa de acerto adjacente. No entanto, essa melhoria implica um impacto significativo em todas as outras métricas. Além disso, esta abordagem requer dois passos de treino, o que implica um consumo adicional de recursos em comparação com as restantes abordagens. Como tal, a sua aplicação não é vantajosa. Por outro lado, a adaptação baseada na média ponderada da distribuição de probabilidade não só aproxima o desempenho do modelo de classificação ao desempenho do modelo de regressão quanto aos valores de REQM e da taxa de acerto adjacente, como ainda melhora o seu desempenho quanto à taxa de acerto e da medida F_1. Isto sugere que ponderar a probabilidade atribuída a cada nível em vez de simplesmente selecionar aquele com a maior probabilidade é uma abordagem adequada para considerar a natureza ordinal dos níveis.

Para podermos fazer uma análise mais aprofundada do desempenho das diferentes abordagens, a Tabela 7 mostra as matrizes de confusão das execuções com maior taxa de acerto das abordagens de classificação e regressão e da adaptação baseada na média ponderada da distribuição de probabilidade das classes. Tal como observado na comparação entre as versões ajustadas de vários modelos pré-treinados, o nível B1 é aquele para o qual a sensibilidade é maior, o que pode sugerir um certo viés para a sua previsão. No entanto, este é também um dos níveis com maior precisão,

Classificação

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	73	16	3	0	0
	A2	22	133	2	0	0
	B1	3	10	216	6	5
	B2	0	0	14	28	7
	C1	0	2	13	4	41

Regressão

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	72	19	1	0	0
	A2	28	125	4	0	0
	B1	0	9	220	11	0
	B2	0	0	16	31	2
	C1	0	0	11	12	37

Média *Softmax*

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	66	22	3	1	0
	A2	18	137	2	0	0
	B1	0	7	223	3	7
	B2	0	1	11	26	11
	C1	0	2	8	8	42

Tabela 7: Matrizes de confusão dos modelos de classificação e regressão com maior taxa de acerto no cenário de validação cruzada: classificação (82,11%), regressão (81,10%) e média *softmax* (82,61%).

superado apenas pelo nível C1 na abordagem de regressão. Do mesmo modo, mantêm-se as dificuldades em distinguir entre si os níveis A. Nesse sentido, enquanto a abordagem de regressão tem mais tendência em prever exemplos do nível A2 como sendo do nível A1, a adaptação do modelo de classificação tem a tendência inversa. Além disso, considerando os níveis mais avançados, podemos ver que a abordagem de regressão parece ter um viés para os níveis B, enquanto a abordagem de classificação e a sua adaptação têm alguma tendência em sobre-estimar a dificuldade de alguns textos desses níveis. Ainda assim, um problema comum a todas as abordagens é a classificação de um conjunto de textos de nível C1 como sendo do nível B1. Tendo em conta que o nível dos exemplos de treino é inferido a partir do nível do respetivo exame, esta divergência na classificação pode significar que esses textos não são realmente de nível C1, sendo a dificuldade introduzida pelo exercício a realizar.

6.2. Generalização

A Tabela 8 mostra os resultados obtidos quando os modelos treinados no cenário de validação cruzada são aplicados no conjunto de teste. Podemos ver que, em geral, o desempenho dos modelos é superior quando é considerado o nível do exame em vez da anotação da complexidade dos textos feita por peritos. No entanto, isto era expectável, uma vez que os modelos foram treinados em exemplos cujo nível foi aproximado com base no nível do exame. Por outro lado, mesmo considerando o nível do exame, o melhor desempenho é de apenas apenas de 46,88% em taxa de acerto e de 58,39% na medida F_1, o que revela uma fraca capacidade de generalização. Se considerarmos execuções individuais, a abordagem de regressão demonstra uma maior capacidade de generalização, tal como observado por Heilman et al. (2008). Ainda assim, a taxa de acerto máxima continua a ser de apenas 50%. Logo, é importante analisar os resultados com mais detalhe para perceber a causa desta quebra de desempenho em comparação com o cenário de validação cruzada e encontrar soluções para a mitigar.

A Tabela 9 apresenta as matrizes de confusão das melhores execuções das várias abordagens quando aplicadas ao conjunto de teste, considerando o nível do exame. Podemos ver que as principais diferenças entre as abordagens que foram observadas no cenário de validação cruzada também podem ser observadas neste caso. Contudo, também podemos ver que todos os modelos preveem vários exemplos dos níveis A como sendo do nível B1. Sem mais informações, poderíamos assumir que os modelos estão enviesados para prever o nível predominante nos dados de treino. Porém, ao inspecionar esses exemplos, descobrimos que correspondem aos textos curtos, como os mostrados na Tabela 3, que são exclusivos dos exames modelo dos níveis A. O desempenho ligeiramente superior da abordagem de regressão deve-se ao facto de ser a única capaz de classificar corretamente alguns destes textos. A classificação da maioria destes textos como B1 pode ser explicada pelo facto de os textos mais curtos nos dados de treino pertencerem a esse nível, embora sejam significativamente mais longos. Como tal, acreditamos que a incapacidade dos modelos para generalizar o seu desempenho para este tipo de texto pode ser superada pela inclusão de tipos de texto mais diversificados nos dados de treino.

A Tabela 10 mostra os resultados se os textos curtos problemáticos não forem considerados. Neste caso, já é possível observar um desempenho bastante mais próximo, apesar de ainda inferior, ao observado no cenário de validação cruzada. A exceção é a abordagem baseada em características usada como linha de base, que continua a revelar dificuldades de generalização. Isto deve-se a um sobreajustamento aos dados de treino, que potencialmente poderá ser mitigado através do uso de apenas um subconjunto das características. No entanto, deixamos esse estudo para trabalho futuro.

Tal como no cenário de validação cruzada, neste caso, a abordagem de regressão é a que leva à maior taxa de acerto adjacente, atingindo um resultado perfeito. Por outro lado, a abordagem de classificação é a com melhor

	Exame			Anotadores		
Abordagem	**TA**	**TAA**	**F_1**	**TA**	**TAA**	**F_1**
Linha de Base	34,38±0,00	75,00±0,00	39,30±0,00	25,00±0,00	81,25±0,00	29,77±0,00
Classificação	**46,88±0,00**	78,12±0,00	**58,39±0,00**	28,12±0,00	**84,38±0,00**	34,80±0,00
Regressão	**46,88±3,12**	**86,46±1,80**	52,40±2,66	**46,88±3,12**	**84,38±0,00**	**52,62±1,74**
Média *Softmax*	45,83±1,80	78,12±0,00	56,16±3,85	29,17±1,80	**84,38±0,00**	36,76±3,39

Tabela 8: Resultados obtidos no conjunto de teste pelas diferentes abordagens baseadas no ajuste de modelos, considerando os níveis dos exames e as anotações dos peritos.

Linha de Base

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	2	1	5	0	0
	A2	2	1	8	0	1
	B1	1	0	4	0	0
	B2	0	0	0	2	1
	C1	0	0	1	1	2

Classificação

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	3	0	5	0	0
	A2	2	2	8	0	0
	B1	1	0	4	0	0
	B2	0	0	0	3	0
	C1	0	0	1	0	3

Regressão

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	3	1	4	0	0
	A2	1	4	7	0	0
	B1	0	1	4	0	0
	B2	0	0	0	3	0
	C1	0	0	0	2	2

Média *Softmax*

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	3	0	5	0	0
	A2	2	2	8	0	0
	B1	1	0	4	0	0
	B2	0	0	0	3	0
	C1	0	0	1	0	3

Tabela 9: Matrizes de confusão das execuções de cada abordagem com maior taxa de acerto no conjunto de teste considerando o nível do exame: linha de base (34,38%), classificação (46,88%), regressão (50,00%) e média *softmax* (46,88%).

desempenho em termos de taxa de acerto e medida F_1. Além disso, é também a mais estável, não revelando qualquer variação de desempenho entre execuções. Pelo contrário, o desempenho inferior revelado pela adaptação com base na média da distribuição de probabilidades deve-se ao fraco desempenho numa das execuções. Contudo, no nosso estudo com mais execuções (Ribeiro et al., 2024b), observámos que, em média, é possível obter um desempenho superior usando esta adaptação.

Além da dificuldade em classificar os textos curtos, existem mais alguns erros comuns a todas as abordagens. Por exemplo, existe um diálogo extraído do exame modelo do nível A2 que é classificado como sendo de nível A1 por todos os modelos. Existe também a tendência de classificar alguns textos do exame do nível C1 como sendo dos níveis B. No entanto, é importante relembrar que a classificação do nível de complexidade é uma tarefa subjetiva e difícil, até para humanos (Branco et al., 2014a; Curto, 2014). Na verdade, todos os textos que foram agora referidos, foram classificados pelos peritos de forma semelhante aos modelos. Isto sugere que os modelos têm alguma capacidade de identificar exemplos cujo nível está desfasado do nível do exame que os contém.

Voltando a analisar as Tabelas 8 e 10, mas focando agora nos resultados em que são consideradas as anotações dos peritos, podemos ver que, tal como referido anteriormente, com exceção da abordagem de regressão, o desempenho dos modelos em termos de taxa de acerto e medida F_1 é bastante inferior ao atingido quando é considerado o nível do exame. A abordagem de regressão é uma exceção por duas razões. Primeiro, é a abordagem com maior variação entre execuções, permitindo-lhe ter simultaneamente a execução com melhor desempenho considerando o nível do exame e a execução com melhor desempenho considerando as anotações dos peritos, com uma taxa de acerto de 50% em ambos os casos. Segundo, esta abordagem evita prever os níveis extremos, especialmente o nível C1. Os peritos revelaram um comportamento semelhante. Como tal, mesmo as execuções com menor desempenho aproximam-se mais das anotações dos peritos do que as da abordagem de classificação, que tem mais tendência para prever esses níveis.

	Exame			Anotadores		
Abordagem	**TA**	**TAA**	**F_1**	**TA**	**TAA**	**F_1**
Linha de Base	57,89±0,00	84,21±0,00	56,00±0,00	42,10±0,00	89,47±0,00	40,32±0,00
Classificação	**78,95±0,00**	89,47±0,00	**79,81±0,00**	47,37±0,00	**94,74±0,00**	46,60±0,00
Regressão	75,44±2,48	**100,00±0,00**	74,60±2,58	**71,93±2,48**	**94,74±0,00**	**70,22±2,51**
Média *Softmax*	77,19±2,48	89,47±0,00	77,59±3,14	49,12±2,48	**94,74±0,00**	48,56±2,77

Tabela 10: Resultados obtidos no conjunto de teste pelas diferentes abordagens baseadas no ajuste de modelos quando os textos curtos extraídos dos exames modelo dos níveis A não são considerados.

Linha de Base

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	1	1	4	0	0
	A2	3	1	10	0	0
	B1	1	0	2	0	1
	B2	0	0	2	3	2
	C1	0	0	0	0	1

Classificação

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	2	0	4	0	0
	A2	3	1	10	0	0
	B1	1	1	2	0	0
	B2	0	0	2	3	2
	C1	0	0	0	0	1

Regressão

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	2	0	4	0	0
	A2	0	6	8	0	0
	B1	1	1	2	0	0
	B2	0	0	1	5	1
	C1	0	0	0	0	1

Média *Softmax*

		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	2	0	4	0	0
	A2	3	1	10	0	0
	B1	1	1	2	0	0
	B2	0	0	2	4	1
	C1	0	0	0	0	1

Tabela 11: Matrizes de confusão das execuções de cada abordagem com maior taxa de acerto no conjunto de teste considerando as anotações dos peritos: linha de base (25,00%), classificação (28,12%), regressão (50,00%) e média *softmax* (31,25%).

Estas tendências podem ser confirmadas nas matrizes de confusão apresentadas na Tabela 11.

Por fim, na Tabela 12, podemos ver o desempenho dos modelos considerando apenas os textos em que existe concordância entre o nível do exame e a anotação dos peritos. Verifica-se que o desempenho de todas as abordagens neste subconjunto é superior ao observado no conjunto de teste completo, independentemente da anotação considerada. No entanto, grande parte desta melhoria no desempenho se deve ao facto de cerca de metade dos textos curtos não estarem a ser considerados. Ainda assim, é interessante constatar que a abordagem de regressão é aquela com melhor desempenho em todas as métricas, falhando apenas a classificação de alguns dos restantes textos curtos. Por outro lado, além dos textos curtos, a abordagem de classificação e a sua adaptação classificam erradamente, como sendo do nível A1, o exemplo do nível A2 mostrado no Apêndice A.2.

Em geral, considerando a quantidade reduzida de dados disponível para treinar os modelos e a inferência do seu nível de complexidade a partir do nível do exame, a abordagem de regressão parece ser aquela com maior capacidade de generalização, uma vez que tem menos tendência para sobreajustamento. Assim, se se dispusesse de um conjunto de treino maior, mais diversificado e mais representativo, esperar-se-ia que a abordagem de classificação e, em particular, a sua adaptação com base na média ponderada da distribuição de probabilidades, atingisse um desempenho superior.

6.3. Modelos Baseados em Instruções

A Tabela 13 mostra os resultados obtidos usando modelos baseados em instruções num cenário *zero-shot*. Em primeiro lugar, é preciso constatar que é difícil encontrar padrões interessantes nos resultados. Ainda assim, podemos ver que o melhor desempenho ao nível de todas as métricas é obtido por modelos da família GPT. O modelo e instrução que levam ao melhor desempenho variam consoante a métrica e se é considerado o nível do exame ou a anotação dos peritos.

Considerando o nível do exame, o melhor desempenho em termos de taxa de acerto (55,21%) e medida F_1 (56,63%) é atingido pelo modelo

Abordagem	**TA**	**TAA**	**F_1**
Linha de Base	40,00±0,00	**93,33±0,00**	41,78±0,00
Classificação	55,33±0,00	**93,33±0,00**	63,00±0,00
Regressão	**64,44±3,14**	**93,33±0,00**	**72,89±1,89**
Média *Softmax*	55,33±0,00	**93,33±0,00**	63,00±0,00

Tabela 12: Resultados obtidos no conjunto de teste pelas diferentes abordagens baseadas no ajuste de modelos, considerando apenas os textos em que a anotação dos peritos coincide com o nível do exame.

		Exame			**Anotadores**		
Modelo	**I**	**TA**	**TAA**	**F_1**	**TA**	**TAA**	**F_1**
GPT 3.5 Turbo	A	**55,21±1,47**	92,71±1,47	**56,63±0,96**	37,50±2,55	95,83±1,47	39,97±2,93
	B	44,79±5,31	89,58±1,47	41,21±6,05	22,92±2,95	89,58±1,47	23,38±3,66
GPT 4 Turbo	A	30,21±1,47	**93,75±0,00**	28,19±1,78	**53,12±0,00**	95,83±1,47	40,35±0,11
	B	36,46±1,47	91,67±1,47	34,21±3,76	50,00±2,55	94,79±1,47	**48,70±8,94**
GPT 4o	A	44,79±1,47	92,71±1,47	47,26±1,28	38,54±1,47	96,88±0,00	38,82±0,96
	B	34,38±0,00	90,62±0,00	33,84±1,37	45,83±2,95	93,75±0,00	41,56±2,87
GPT 4o Mini	A	34,38±0,00	90,62±0,00	37,22±0,00	43,75±0,00	**100,00±0,00**	43,27±0,00
	B	37,50±0,00	90,62±0,00	42,54±0,00	46,88±0,00	**100,00±0,00**	47,58±0,00
Llama 3 8B	A	31,25±0,00	87,50±0,00	26,78±0,00	43,75±0,00	87,50±0,00	34,67±0,00
	B	34,38±0,00	87,50±0,00	26,91±0,00	50,00±0,00	90,62±0,00	38,94±0,00
Mixtral 8x7B	A	43,75±0,00	84,38±0,00	29,52±0,00	31,25±0,00	96,88±0,00	24,24±0,00
	B	46,88±0,00	81,25±0,00	31,75±0,00	31,25±0,00	90,62±0,00	21,50±0,00

Tabela 13: Resultados obtidos no conjunto de teste usando modelos baseados em instruções. A coluna I distingue entre as duas instruções usadas no nosso estudo.

GPT 3.5 Turbo, o mais antigo de todos os explorados, quando emparelhado com a instrução A, que foca diretamente no nível de complexidade do texto. Por outro lado, considerando a anotação dos peritos, o melhor desempenho é atingido pelo modelo GPT 4 Turbo. No entanto, enquanto a maior taxa de acerto (53,12%) é atingida usando a instrução A, a maior medida F_1 (48,70%) é atingida usando a instrução B, que foca no nível de proficiência necessário para compreender o texto. Já a maior taxa de acerto adjacente é atingida pelo modelo GPT 4 Turbo quando é considerado o nível do exame e pelo modelo GPT 4o Mini quando são consideradas as anotações dos peritos. Contudo, em ambos os casos, este desempenho surge às custas de uma quebra elevada em termos das restantes métricas, especialmente quando considerando o nível do exame.

Considerando os resultados referidos anteriormente, poderíamos ser levados a assumir que os modelos baseados em instruções têm um desempenho melhor quando é considerado o nível do exame, pois podem potencialmente ter visto os exames modelo durante o seu treino. Porém, analisando a Tabela 13, podemos ver que o modelo GPT 3.5 Turbo é uma exceção e que, na maioria dos casos, o desempenho é melhor quando são consideradas as anotações dos peritos.

Em comparação com os resultados dos modelos de base ajustados para a tarefa, o desempenho dos melhores modelos baseados em instruções é superior em termos da taxa de acerto e da taxa de acerto adjacente, mas inferior ao nível da medida F_1. Analisando as matrizes de confusão na Tabela 14, podemos ver que o aumento da taxa de acerto se deve ao facto de os modelos não estarem enviesados para classificar os textos curtos como sendo do nível B1. Por outro lado, a medida F_1 inferior deve-se a uma maior confusão entre os níveis A e à incapacidade de o modelo GPT 4 Turbo prever o nível C1. Analisando a confusão entre os níveis A com mais detalhe, descobrimos que esta está, mais uma vez, relacionada com os textos curtos, já que os modelos baseados em instruções têm tendência a classificar a maioria desses textos como sendo do nível A1.

Relativamente aos modelos disponíveis publicamente, podemos ver que, na maioria dos casos, atingem melhores resultados quando emparelhados com a instrução B. Além disso, enquanto o modelo da família Llama tem melhor desempenho quando são consideradas as anotações dos peritos, o modelo da família Mistral tem melhor desempenho quando é considerado o nível do exame. Em ambos os casos, estes modelos ficam em segundo lugar quanto à taxa de acerto e, ainda assim, distantes do melhor desempenho.

		Exame				
		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	5	2	1	0	0
	A2	5	5	1	1	0
	B1	0	0	3	2	0
	B2	0	0	0	2	1
	C1	0	0	0	1	3

		Anotadores				
		Previsto				
		A1	**A2**	**B1**	**B2**	**C1**
Real	**A1**	4	2	0	0	0
	A2	4	6	3	1	0
	B1	0	1	1	2	0
	B2	0	0	1	6	0
	C1	0	0	0	1	0

Tabela 14: Matrizes de confusão das melhores execuções dos modelos baseados em instruções com maior taxa de acerto no conjunto de teste, considerando os níveis dos exames (GPT 3.5 Turbo, Instrução A, 56,25%) e as anotações dos peritos (GPT 4 Turbo, Instrução A, 53,12%).

Alguns comentários mais genéricos sobre os modelos: ao contrário dos modelos disponíveis publicamente, mesmo otimizando os hiperparâmetros para esse efeito, é impossível garantir a reprodutibilidade dos resultados dos modelos da família GPT; os modelos das famílias GPT e Mistral têm alguma dificuldade em seguir a instrução de responder apenas com o nível, independentemente da forma como essa instrução é dada; o modelo da família Mistral inclui dois níveis na resposta em múltiplas ocasiões.

Tendo em conta o desempenho do modelo GPT 3.5 Turbo em conjunto com a instrução A no conjunto de teste quando considerado o nível do exame, decidimos aplicá-lo também ao conjunto de treino, de forma a comparar os resultados com o do cenário de validação cruzada. Contudo, neste caso, a taxa de acerto média foi de apenas 38,18% e a medida F_1 de apenas 33,94%.

Em geral, apesar de aparentarem ter menos enviesamentos que os modelos ajustados especificamente para a tarefa e de serem capazes de lidar com tipos de texto mais diversos, a maioria dos resultados obtidos neste estudo sugerem que os modelos baseados em instruções não têm, de base, capacidade para identificar adequadamente o nível de complexidade dos textos. No entanto, é possível que desenvolvam essa capacidade se as instruções incluírem exemplos adequados. Como tal, no futuro, é importante explorar o desempenho destes modelos num cenário *few-shot*.

7. Conclusão

Neste estudo, explorámos a avaliação automática do nível de complexidade de textos em português europeu. Para isso, começámos por complementar o conjunto de textos extraídos de exames para avaliação do nível de proficiência usado em estudos anteriores com um novo conjunto de teste disponível publicamente e anotado não só por inferência a partir do nível do exame, como também por peritos. Em seguida, analisámos o desempenho de múltiplos modelos de base pré-treinados quando ajustados para desempenhar esta tarefa. Além disso, explorámos não só o uso de abordagens de classificação, como também de regressão, numa tentativa de ter em conta a natureza ordinal dos níveis de complexidade definidos pelo CEFR. Por fim, fizemos uma análise preliminar da capacidade de modelos baseados em instruções desempenharem esta tarefa num cenário *zero-shot* (sem exemplos).

As nossas experiências num cenário de validação cruzada mostraram que, considerando a quantidade reduzida de dados de treino, o melhor desempenho pode ser atingido através do ajuste da versão mais pequena do modelo Albertina PT-PT. Além disso, uma adaptação do modelo de classificação, baseada na previsão com base na média ponderada da distribuição de probabilidades em vez da simples seleção da classe mais provável, leva a previsões mais robustas e, em geral, a um melhor desempenho.

Contudo, também devido à escassez de dados de treino, estes modelos não generalizam bem para diferentes tipos de texto nem para as anotações dos peritos, ficando sobreajustados aos níveis inferidos a partir do nível do exame. Um modelo baseado em regressão pura tem menos tendência a ficar sobreajustado e, por isso, generaliza melhor para tipos de texto não observados durante o treino. No entanto, em contrapartida, o desempenho global é mais baixo. Como tal, tal como para muitas outras tarefas de PLN em línguas com poucos recursos, é importante obter um conjunto de dados de treino maior, mais diversificado e mais representativo, de forma a treinar melhores modelos.

Relativamente ao uso de modelos baseados em instruções, apesar de aparentarem ter menos enviesamentos do que os modelos ajustados especificamente para a tarefa e de serem capazes de lidar com tipos de texto mais diversos, demonstraram também um comportamento bastante errático e de difícil análise, sendo o melhor desempenho em relação aos modelos ajustados atingido apenas em casos muito específicos.

Como tal, argumentamos que a maioria destes modelos não tem, de base, capacidade para identificar adequadamente o nível de complexidade dos textos. Contudo, é possível que venham a desenvolver essa capacidade se as instruções incluírem exemplos adequados. Como tal, como trabalho futuro, pretendemos explorar o desempenho destes modelos em cenários *few-shot*, usando diferentes abordagens para selecionar os exemplos.

Caso a abordagem *few-shot* não resulte, para mitigar os problemas criados pela escassez de dados anotados, pretendemos explorar o uso de textos na variedade brasileira da língua durante o treino e averiguar se as vantagens do volume de dados de treino adicional superam os problemas introduzidos pelas diferenças entre as duas variedades da língua. Esta abordagem pode também ser generalizada para o uso de dados anotados noutras línguas em combinação com modelos de base multilingues.

Uma outra abordagem a explorar no futuro é o uso de modelos híbridos que combinam características linguísticas com modelos de língua, que já revelaram ser capazes de atingir um desempenho superior à soma das partes em estudos anteriores sobre a classificação automática do nível de complexidade de textos (Correia & Mendes, 2021; Wilkens et al., 2024).

Por fim, considerando a subjetividade da avaliação do nível de inteligibilidade ou de complexidade de textos e as suas potenciais aplicações, é importante fazer um esforço para o desenvolvimento de modelos interpretáveis para esta tarefa, de modo a entender porque é que um texto é classificado como sendo de um determinado nível e como ele pode ser alterado para ser mais adequado ao nível de proficiência do público-alvo.

Agradecimentos

Este trabalho foi financiado por fundos nacionais portugueses através da Fundação para a Ciência e a Tecnologia (FCT) (Referência: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) e pela Comissão Europeia (Projeto: iRead4Skills, Número da subvenção: 1010094837, Tópico: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

Gostaríamos de agradecer ao Camões, I.P., por nos ter dado acesso aos textos usados nos seus exames e a permissão para os usar para treinar e avaliar os nossos modelos. Gostaríamos também de agradecer aos peritos que se disponibilizaram para anotar os textos do conjunto de teste.

Referências

Akef, Soroosh, Amália Mendes, Detmar Meurers & Patrick Rebuschat. 2024. Investigating the generalizability of Portuguese readability assessment models trained using linguistic complexity features. Em *International Conference on Computational Processing of Portuguese (PROPOR)*, 332–341.

Aluisio, Sandra, Lucia Specia, Caroline Gasperin & Carolina Scarton. 2010. Readability assessment for text simplification. Em *NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, 1–9.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson et al. 2021. On the opportunities and risks of foundation models. arXiv [cs.LG/cs.AI/cs.CY]. 10.48550/arXiv.2108.07258

Branco, António, João Rodrigues, Francisco Costa, João Silva & Rui Vaz. 2014a. Assessing automatic text classification for interactive language learning. Em *International Conference on Information Society (i-Society)*, 70–78. 10.1109/i-society.2014.7009014

Branco, António, João Rodrigues, Francisco Costa, João Silva & Rui Vaz. 2014b. Rolling out text categorization for language learning assessment supported by language technology. Em *International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, 256–261. 10.1007/978-3-319-09761-9_29

Cha, Miriam, Youngjune Gwon & H.T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. Em *Conference on Information and Knowledge Management (CIKM)*, 2003–2006. 10.1145/3132847.3133104

Chen, Banghao, Zhaofeng Zhang, Nicolas Langrené & Shengxin Zhu. 2023. Unleashing the potential of prompt enginee-

ring in large language models: a comprehensive review. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2310.14735

Correia, João & Rui Mendes. 2021. Neural complexity assessment: A deep learning approach to readability classification for European Portuguese corpora. Em *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*, 300–311. 10.1007/978-3-030-91608-4_30

Council of Europe. 2001. *Common european framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press.

Crossley, Scott A., Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara & Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes* 54(5–6). 340–359. 10.1080/0163853x.2017.1296264

Curto, Pedro. 2014. *Classificador de Textos para o Ensino de Português como Segunda Língua*: Instituto Superior Técnico, Universidade de Lisboa. Tese de Mestrado.

Curto, Pedro, Nuno Mamede & Jorge Baptista. 2015. Automatic text difficulty classifier. Em *International Conference on Computer Supported Education (CSEDU)*, 36–44. 10.5220/0005428300360044

Devlin, Jacob, Ming-Wei Chang, Lee Kenton & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Direção de Serviços de Língua e Cultura. 2017. *Referencial Camões PLE.* Camões, Instituto da Cooperação e da Língua I.P.

DuBay, William H. 2004. *The principles of readability.* Impact Information.

Filighera, Anna, Tim Steuer & Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. Em *European Conference on Technology Enhanced Learning (EC-TEL)*, 335–348. 10.1007/978-3-030-29736-7_25

Forti, Luciana, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci & Stefania Spina. 2020. MALT-IT2: a new resource to measure text difficulty in light of CEFR levels for Italian L2 learning. Em *Language Resources and Evaluation Conference (LREC)*, 7204–7211.

François, Thomas & Cédrick Fairon. 2012. An "AI readability" formula for French as a foreign language. Em *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 466–477.

François, Thomas, Adeline Müller, Eva Rolin & Magali Norré. 2020. AMesure: A web platform to assist the clear writing of administrative texts. Em *Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, 1–7. 10.18653/v1/2020.aacl-demo.1

Friedman, Jerome, Trevor Hastie & Robert Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics* 28(2). 337–407. 10.1214/aos/1016218223

Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29(5). 10.1214/aos/1013203451

Giray, Louie. 2023. Prompt engineering with ChatGPT: a guide for academic writers. *Annals of Biomedical Engineering* 51(12). 2629–2633. 10.1007/s10439-023-03272-4

Grosso, Maria José, António Soares, Fernanda de Sousa & José Pascoal. 2011. QuaREPE: Quadro de referência para o ensino Português no estrangeiro – documento orientador. Relatório técnico. Direção-Geral da Educação (DGE).

Guillou, Pierre. 2020. Faster than training from scratch: Fine-tuning the English GPT-2 in any language with Hugging Face and FastAI v2 (practical case with Portuguese). Medium.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva & Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. Em *International Conference on Learning Representations (ICLR)*,

Heilman, Michael, Kevyn Collins-Thompson & Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. Em *Workshop on Innovative use of NLP for Building Educational Applications (BEA)*, 71–79.

Hernandez, Nicolas, Nabil Oulbaz & Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in French. Em *Proceedings of the Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 54–61.

Hulstijn, Jan H. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91(4). 663–667. 10.1111/j.1540-4781.2007.00627_5.x

Hulstijn, Jan H. 2011. Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly* 8(3). 229–249. 10.1080/15434303.2011.565844

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix & William El Sayed. 2023. Mistral 7B. arXiv [cs.CL/cs.AI/cs.LG]. 10.48550/arXiv.2310.06825

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix & William El Sayed. 2024. Mixtral of experts. arXiv [cs.LG/cs.CS]. 10.48550/arXiv.2401.04088

Jönsson, Simon, Evelina Rennes, Johan Falkenjack & Arne Jönsson. 2018. A component based approach to measuring text complexity. Em *7th Swedish Language Technology Conference (SLTC)*, 58–61.

Karpov, Nikolay, Julia Baranova & Fedor Vitugin. 2014. Single-sentence readability prediction in Russian. Em *International Conference on Analysis of Images, Social Networks and Texts (AIST)*, 91–100. 10.1007/978-3-319-12580-0_9

Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers & Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Relatório técnico. Institute for Simulation and Training, University of Central Florida. 10.21236/ada006655

Leal, Sidney Evaldo & Sandra Maria Aluísio. 2024. Complexidade textual e suas tarefas relacionadas. Em Helena. M. Caseli & Maria G. V. Nunes (eds.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chap. 23. BPLN 2nd edn.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2023. NILC-Metrix: Assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation* 73–110. 10.1007/s10579-023-09693-w

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: a robustly optimized BERT pretraining approach. arXiv [cs.CL]. 10.48550/arXiv.1907.11692

Llama Team. 2024. The Llama 3 herd of models. arXiv [cs.AI/cs.CL/cs.CV]. 10.48550/arXiv.2407.21783

Martinc, Matej, Senja Pollak & Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics* 47(1). 141–179. 10.1162/coli_a_00398

Marujo, Luís, José Lopes, Nuno Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista & Céu Viana. 2009. Porting REAP to European Portuguese. Em *International Workshop on Speech and Language Technology in Education (SLaTE)*, 69–72.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42(2). 109–127. 10.1111/j.2517-6161.1980.tb01109.x

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and

their compositionality. Em *Advances in Neural Information Processing Systems (NIPS)*, 3111–3119.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral & Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. Em *Annual Meeting of the Association for Computational Linguistics (ACL)*, 3470–3487. 10.18653/v1/2022.acl-long.244

Mohtaj, Salar, Babak Naderi & Sebastian Möller. 2022. Overview of the GermEval 2022 shared task on text complexity assessment of German text. Em *GermEval Workshop on Text Complexity Assessment of German Text*, 1–9.

Nadeem, Farah & Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. Em *Workshop on Innovative Use of NLP for Building Educational Applications*, 45–55. 10.18653/v1/W18-0505

North, Kai, Marcos Zampieri & Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys* 55(9). 1–42. 10.1145/3557885

OpenAI. 2023. ChatGPT. Online.

OpenAI. 2024. GPT-4 technical report. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2303.08774

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike & Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Em *Advances in Neural Information Processing Systems (NIPS)*, 27730–27744.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(84). 2825–2830.

Pilán, Ildikó & Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. Em *Workshop on Linguistic Complexity and Natural Language Processing*, 49–58.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI Blog.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Reynolds, Robert. 2016. Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. Em *Workshop on Innovative Use of NLP for Building Educational Applications*, 289–300. 10.18653/v1/W16-0534

Ribeiro, Eugénio, Nuno Mamede & Jorge Baptista. 2024a. Automatic text readability assessment in European Portuguese. Em *International Conference on Computational Processing of Portuguese (PROPOR)*, 97–107.

Ribeiro, Eugénio, Nuno Mamede & Jorge Baptista. 2024b. Text readability assessment in European Portuguese: A comparison of classification and regression approaches. Em *International Conference on Computational Processing of Portuguese (PROPOR)*, 551–557.

Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso & Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertina PT-*. Em *Portuguese Conference on Artificial Intelligence (EPIA)*, 441–453. 10.1007/978-3-031-49008-8_35

Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv [cs.CL]. 10.48550/arXiv.1910.01108

Santos, Rodrigo, João Rodrigues, António Branco & Rui Vaz. 2021. Neural text categorization with transformers for learning Portuguese as a second language. Em *Portuguese Conference on Artificial Intelligence (EPIA)*, 715–726. 10.1007/978-3-030-86230-5_56

Santos, Rodrigo, João Rodrigues, Luís Gomes, João Ricardo Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório & Bernardo Leite. 2024. Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family. Em *Proceedings of the Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL)*, 105–114.

Santucci, Valentino, Filippo Santarelli, Luciana Forti & Stefania Spina. 2020. Automatic classification of text complexity. *Applied Sciences* 10(20). 7285. 10.3390/app10207285

Scarton, Carolina Evaristo & Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: Adaptando as métricas do Coh-Metrix para o Português. *Linguamática* 2(1). 45–61.

Seabold, Skipper & Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. Em *Python in Science Conference (SciPy)*, 92–96. 10.25080/majora-92bf1922-011

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. 10.1007/978-3-030-61377-8_28

Suárez, Pedro Javier Ortiz, Benoît Sagot & Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. Em *Workshop on the Challenges in the Management of Large Corpora (CMLC)*, 9–16. 10.14618/ids-pub-9021

Sung, Yao Ting, Wei Chun Lin, Scott Benjamin Dyson, Kuo En Chang & Yu Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal* 99(2). 371–391. 10.1111/modl.12213

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave & Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. arXiv [cs.CL]. 10.48550/arXiv.2302.13971

Wagner Filho, Jorge A., Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: a new open resource for Brazilian Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Wilkens, Rodrigo, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey & Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. Em *Language Resources and Evaluation Conference (LREC)*, 1217–1233.

Wilkens, Rodrigo, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont & Thomas François. 2024. Exploring hybrid approaches to readability: Experiments on the complementarity between linguistic features and transformers. Em *Findings of the Association for Computational Linguistics*, 2316–2331.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander M. Rush. 2020. HuggingFace's transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Xia, Menglin, Ekaterina Kochmar & Ted Briscoe. 2016. Text readability assessment for second language learners. Em *Workshop on Innovative Use of NLP for Building Educational Applications*, 12–22. 10.18653/v1/W16-0502

Yancey, Kevin, Alice Pintard & Thomas Francois. 2021. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio* 20(2). 229–258. 10.1418/102814

Zago, Ricardo. 2023. BERTugues base (aka "bertugues-base-portuguese-cased"). Hugging Face.

A. Exemplos

Este apêndice mostra um exemplo de cada nível de complexidade, extraído dos exames modelo do Camões, I.P., e selecionado de entre os casos em que existe concordância entre o nível do exame e o atribuído pelos anotadores. Quando disponíveis, as referências às fontes dos textos estão incluídas por uma questão de atribuição. No entanto, estas não são consideradas parte do texto.

A.1. A1

Dois irmãos, o Diogo e o Carlos, vão comprar uma prenda para o pai.

Diogo: Carlos, este CD é bom. Gosto muito deste cantor!
Carlos: Sim, eu também gosto. Mas o pai prefere livros, não achas?
Diogo: Sim, tens razão. Ele está sempre a ler.
Carlos: Pois está! Podemos dar-lhe um livro sobre reis e rainhas.
Diogo: Boa ideia. O pai gosta de História de Portugal.
Carlos: Vê lá este aqui, chama-se Filipa de Lencastre.
Diogo: Esse livro deve ser giro. Acho que o pai vai gostar. Quanto é que custa?
Carlos: Custa 15€. Não é muito caro, pois não?
Diogo: Não, é um bom preço.
Carlos: Vamos então levar este livro? Temos aqui esse dinheiro.
Diogo: Sim, acho que é um bom presente.

A.2. A2

Portugal nasceu no Norte. Foi na região Porto e Norte que os portugueses começaram enquanto povo e nação.

O Porto é a grande porta de entrada e pode ser ponto de partida para uma viagem pela diversidade natural e cultural da região. É conhecido pelo vinho e por arquitetos como Álvaro Siza Vieira e Souto de Moura.

O rio Douro atravessa a região. Entra em Portugal apertado por montanhas do interior para percorrer toda a paisagem onde se cultiva o vinho do Porto.

Turismo de Portugal, https://www.visitportugal.com/pt-pt/destinos/porto-e-norte (texto adaptado)

A.3. B1

Dádivas

Anualmente, no Natal, são oferecidos animais. Muitos deles, comprados a lojas. Enquanto isso, nas associações de animais, existem milhares que esperam uma casa quente e o colo de alguém.

A maior parte já foi abandonada pelo mesmo colo que lhe jurou amor eterno, mas que o largou à primeira dificuldade. As desculpas já as conhecemos: ou porque se vai de férias, ou porque nasceu um bebé, ou porque o animal está doente e não se tem dinheiro para tratá-lo.

O problema está em ver-se um animal como um objeto em vez de se pensar nele como um ser que faz parte de uma família. Os animais criam ligações emocionais, têm memória e sofrem quando magoados física e psicologicamente. Se assim não fosse, como explicaríamos os casos de animais que arriscam a vida para salvar os seus companheiros humanos?

Antes de se ter um animal, deve-se refletir se se estará disposto a enfrentar todas as dificuldades. Um animal não é uma prenda. É uma dádiva que só alguns saberão reconhecer.

Ana Bacalhau, *Notícias Magazine*, 28 dezembro 2014 (texto adaptado)

A.4. B2

Imitar o voo das aves

Há quem veja os voadores como "doidos" do ar, por sentirem um desejo irreprimível de voar. O mais curioso é que esse desejo sempre está presente na humanidade, provavelmente desde o dia em que se reparou no voo dos pássaros. Aquilo que terá começado como uma mera contemplação, transformou-se numa obsessão: imitar as aves.

Embora voar fosse para a maioria das pessoas um sonho impossível, algo que estaria para além da capacidade humana, vários foram os que acreditaram e tudo fizeram para o ver concretizado. Uns ficaram-se por cálculos e arabescos em folhas de papel, outros, mais audazes, lançaram-se no vazio com as engenhocas voadoras que construíram. Desses, muitos pagaram com a própria vida a ousadia, outros colecionaram desastres e ferimentos graves. Afortunadamente, um punhado de visionários conseguiu ser bem-sucedido, inscrevendo para sempre os seus nomes na história da aviação. [...]

Da Grécia Antiga, chega-nos o relato daquela que poderá ter sido a primeira máquina voadora movimentada por meios próprios. Diz-se que foi construída por volta de 400 a.C., pelo ilustre matemático Arquitas de Tarento, e seria um pombo de madeira movido a jatos de vapor, que dizem ter voado cerca de duas centenas de metros.

A poesia árabe exalta os feitos de Abbas ibn Firnas (810–887), que alguns consideram ter sido a primeira pessoa a construir um engenho voador e a voar. [...]

O inventor e artista Leonardo da Vinci, génio do Renascimento, passou anos a decifrar o voo das aves e a conceber planos meticulosos para máquinas voadoras. Porém, como não conseguiu perceber as leis da física associadas ao voo, acabou por deixar o mundo dos vivos sem nunca ter construído as suas máquinas ou experimentado o prazer de voar.

Revista *Super Interessante*, novembro de 2012 (texto adaptado e com supressões)

A.5. C1

A mentira da criatividade

O sentido da palavra criatividade banalizou-se. Hoje não existe anúncio de emprego para empresas ou instituições que não tenha um item a proclamar que se privilegiam pessoas criativas, com capacidade inovadora e ideias fora da caixa.

E se o mercado quer, é sintomático que surjam cada vez mais pessoas com esse perfil. É natural que assim seja. Não existe político ou empresário que nos últimos anos não debite com aparente convicção que inovar é necessário.

Ontem um letreiro avisava-me que tinha acabado de entrar numa 'pastelaria criativa'. Ainda olhei para o meu palmier simples a ver se tinha linhas complexas desenhadas por um designer ou se o sabor era exótico, mas não, era apenas um simples, honesto e saboroso palmier. E suspirei de alívio.

Não me interpretem mal. Ser criativo, inovador e ter ideias fora da caixa, é óptimo. Mas não existem muitas pessoas com esse perfil. E tenho sérias dúvidas que as empresas as desejem. O que temos hoje em dia é a romantização dessas noções e sua apropriação superficial, através da retórica que as envolve.

Por um lado vemos cada vez mais difundido o desígnio de que todos podemos triunfar com uma boa ideia. Na verdade celebramos aqueles que consideramos serem criativos, mas apenas a partir do resultado que produziram. A realidade é esta: a maior parte das pessoas, empresas ou instituições tem dificuldade em lidar com indivíduos realmente criativos.

A prosa poética e o amor pelas histórias vieram da mãe: "Era professora primária e contava muito bem histórias de raiz popular. Mas também histórias de fadas. Eu adorava."

Há exceções? Há, como em tudo. Mas não passam disso.

Vítor Belanciano, in *Público*, 26/04/2015 (texto com supressões)