Automatic Text Readability Assessment in European Portuguese

Eugénio Ribeiro[1] and **Nuno Mamede**[1,2] and **Jorge Baptista**[1,3]
[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract

The automatic assessment of text readability and the classification of texts by 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. In this study, we address the automatic assessment of text readability according to CEFR levels in European Portuguese. For that, we explore the fine-tuning of several foundation models on textual data used for proficiency evaluation purposes. Additionally, we aim at setting the ground for more comparable research on this subject by defining a new publicly available test set. Our experiments show that the best models can achieve around 80% accuracy and 75% macro F1 score. However, they have difficulty in generalizing to different types of text, which reveals the need for additional and more diverse training data.

1 Introduction

Identifying the readability level of a text is relevant across diverse domains, encompassing not only language education but also various language-related industries and many other human activities. In education, assessing the readability level allows educators and curriculum designers to match texts to the learners' abilities, fostering effective language development and personalized learning experiences. Moreover, outside the education domain, readability level classification finds applications in different sectors. For instance, in the banking industry, presenting financial information and policies at an appropriate readability level ensures that clients can understand terms and conditions, enabling well-informed decision-making. Similarly, in healthcare, accessible and understandable medical instructions, consent forms, and patient information materials are crucial for individuals with varying levels of language proficiency. Furthermore, legal information, government communications, user manuals, and many others, benefit from accurately assessing the readability level of written materials, facilitating effective communication, content transparency, and general comprehension.

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) provides a widely recognized framework for classifying language proficiency levels, ranging from A1 (beginner) to C2 (proficient). This framework is typically used to assess the proficiency level of learners of a given language. However, it can also be used from a readability perspective, as a means to identify the proficiency required to understand specific pieces of text. Therefore, by exploring the readability perspective of the CEFR, we can make a significant contribution to enhancing the understanding of text comprehension factors and their far-reaching implications for both education and language-related industries seeking to convey information to learners or clients in a manner that is clear, concise, and easily understood.

Determining the readability level of texts presents its own set of challenges, particularly when working with languages that have limited annotated resources. Annotating large amounts of text data with CEFR levels is a labor-intensive and time-consuming task, often requiring expert domain knowledge. Consequently, the scarcity of labeled data hinders the development of robust and accurate models for automatic readability level classification in multiple languages.

In this study, we address the automatic assessment of text readability according to CEFR levels in European Portuguese. For that, we rely on the recent developments on foundation models for

Portuguese (Rodrigues et al., 2023) and compare the performance of those models with that of previously existing ones when fine-tuned on textual data used for proficiency evaluation purposes by Camões, I.P. [1], the official Portuguese language institute. Additionally, considering that this data is not publicly available and that different subsets of it were used in previous studies on the task (e.g., Branco et al., 2014b; Curto et al., 2015; Santos et al., 2021), we aim at setting the ground for more comparable future research on this subject by defining a new test set based on the model exams that are publicly available on the institute's website.

In the remainder of this document, we start by providing an overview of related work on automatic text readability level assessment, with a focus on European Portuguese in Section 2. Then, in Section 3, we describe our experimental setup, including the dataset, the foundation models, and the methodologies employed for fine-tuning and evaluation. Next, in Section 4, we present and discuss the results of our experiments, including the errors and biases observed for the different models. Finally, in Section 5, we summarize the contributions of this study, discuss its limitations, and provide pointers for future research in the area.

2 Related Work

Readability assessment is a problem that has been widely explored over the years. Traditionally, the problem is addressed by creating readability formulas or indexes based on statistical information and/or domain knowledge (DuBay, 2004; Crossley et al., 2017). Among these, the most widely used are the Flesch Reading Ease Index and the Flesch-Kincaid Grade Level (Kincaid et al., 1975).

However, considering the developments in Machine Learning (ML), and especially in Natural Language Processing (NLP), the research on automatic readability assessment shifted towards following the trends in the NLP area (Graesser et al., 2004; McNamara et al., 2014). This trend was also followed in related tasks, such as lexical complexity assessment (North et al., 2023). Early approaches (and many recent ones for low-resource languages) relied on handcrafted features, such as word frequency, sentence length, and syntactic complexity, combined with traditional machine learning algorithms, such as decision trees and Support Vector Machines (SVMs) (e.g., Aluisio et al., 2010; François and Fairon, 2012; Karpov et al., 2014; Curto et al., 2015; Pilán and Volodina, 2018; Forti et al., 2020; Leal et al., 2023). Then, Deep Learning (DL) approaches relying on pre-trained word embeddings, such as those generated by Word2Vec (Mikolov et al., 2013), emerged (e.g., Cha et al., 2017; Nadeem and Ostendorf, 2018; Filighera et al., 2019). Finally, more recently, research in the area shifted towards the fine-tuning of pre-trained Transformer-based foundation models, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), and RoBERTa (Liu et al., 2019) (e.g., Santos et al., 2021; Yancey et al., 2021; Martinc et al., 2021; Mohtaj et al., 2022).

Similarly to most NLP tasks, a significant part of the research on automatic text readability level assessment focuses on the English language (e.g., Xia et al., 2016; Cha et al., 2017; Nadeem and Ostendorf, 2018; Filighera et al., 2019; Martinc et al., 2021). However, in this case, there are also several studies addressing the problem in other languages, many of which are low-resourced. For instance, there are studies in French (e.g., François and Fairon, 2012; François et al., 2020; Yancey et al., 2021; Wilkens et al., 2022; Hernandez et al., 2022), Chinese (e.g., Sung et al., 2015), German (e.g., Mohtaj et al., 2022), Italian (e.g., Forti et al., 2020; Santucci et al., 2020), Russian (e.g., Karpov et al., 2014; Reynolds, 2016), Swedish (e.g., Jönsson et al., 2018; Pilán and Volodina, 2018), and Slovenian (e.g., Martinc et al., 2021).

Focusing on Portuguese, there are a few studies covering the Brazilian variety of the language (e.g., Scarton and Aluísio, 2010; Aluisio et al., 2010; Leal et al., 2023). However, in this study, we are mainly interested in the European variety. Thus, below, we describe previous studies covering this variety in further detail.

The Portuguese version of the REAP tutoring system (Marujo et al., 2009) included a readability level classifier trained on 5th to 12th-grade textbooks. The model was based on SVMs applied to lexical features, such as statistics of word unigrams, and included additional strategies to capture the ordinal nature of the levels (McCullagh, 1980). Although this model was accurate when applied to school textbooks, its performance significantly decreased when applied to exams of the 6th, 9th, and 12th grades.

LX-CEFR (Branco et al., 2014b) is a tool designed to help language learners and teachers of Portuguese in assessing the CEFR level of a text.

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

It focuses on four different features independently: the Flesch Reading Ease index, the lexical category density in terms of the proportion of nouns, the average word length in number of syllables, and the average sentence length in number of words. A corpus of 114 labeled excerpts extracted from the Portuguese exams performed by Camões, I.P. was used to compute the correlation between these features and the readability level. A subsequent study (Branco et al., 2014a) focused on the re-evaluation of the tool by human experts, as well as the re-annotation of the texts by multiple language instructors. Regarding the latter, the inter-annotator agreement was of just 0.17, which reveals the difficulty and subjectivity of the task.

Curto et al. (2015) explored the use of several traditional ML algorithms for the task. The algorithms were applied to 52 features split into 5 different groups: Part-of-Speech (POS), chunks, sentences and words, verbs, averages and frequencies, and extras. The experiments were performed on an extended version of the dataset used in the context of LX-CEFR containing 237 excerpts. The highest performance was achieved using LogitBoost (Friedman et al., 2000). Additionally, similarly to what was observed by Branco et al. (2014a), a re-annotation of this extended version of the dataset by two groups of multiple experts revealed low inter-annotator agreements of 0.188 and 0.164 (Curto, 2014).

Finally, Santos et al. (2021) explored the use of two neural models for the task. More specifically, they fine-tuned Portuguese versions of the GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) models on multiple variants of the dataset of Camões, I.P. exams to compare the performance not only between the two foundation models, but also with that of previous approaches to the task. Overall, on the larger versions of the dataset, including a new one with 500 excerpts, the fine-tuned GPT-2 model achieved the highest performance. Our study builds on this one by assessing the performance of several additional foundation models and by performing a deeper analysis of their performance, with a focus on the errors and their causes.

3 Experimental Setup

In this section, we describe our experimental setup. We start by describing the dataset used in our experiments in Section 3.1. Then, in Section 3.2, we list the multiple foundation models used in our study. In Section 3.3, we describe the methodology used for fine-tuning those models and evaluate their performance on the task. Finally, in Section 3.4, we provide implementation details that enable the future reproduction of our experiments.

	A1	**A2**	**B1**	**B2**	**C1**	**Total**
Train	92	157	240	49	60	598
Test	8	12	5	3	4	32

Table 1: Distribution of the texts in the dataset of Camões, I.P. exams across CEFR levels.

3.1 Dataset

Similarly to the previous studies on automatic text readability assessment in European Portuguese discussed in Section 2, our dataset is comprised of texts extracted from the Portuguese exams performed by Camões, I.P., the official Portuguese language institute. The texts cover the CEFR levels A1 to C1, as defined in the Portuguese version of the framework (Grosso et al., 2011; Direção de Serviços de Língua e Cultura, Camões, I.P., 2017). Considering that these texts are used for evaluation purposes and can be reused over time, they are not publicly available. This makes it hard for researchers who have no access to the texts to perform research on the task. Furthermore, the number of annotated texts increases over time and there is no standard partitioning of the data. This led to multiple different versions of the dataset being used in the previous studies, which makes it difficult to compare the existing approaches. However, there is a set of model exams (one for each level) that is publicly available on the institute's website. Thus, we propose to extract the texts used for reading comprehension in those exams and use them as a test set. This way, evaluation can be standardized in the future and researchers without access to the private exams can still at least evaluate their approaches on this set.

Table 1 shows the distribution of the texts across CEFR levels. At the time of this study, there were 598 texts available from the private exams. We can see that there is a bias towards the middle (B1) level and fewer examples of the advanced levels. Furthermore, considering that some texts are reused over time, some of the examples consist of small variations of the same text.

The test set extracted from the publicly available

A1	*É favor não jogar à bola no interior da escola.* Please do not play football inside the school.
A1	*É obrigatório desligar o computador antes de sair da sala.* It is mandatory to turn off the computer before leaving the room.
A2	*Lamentamos mas não é possível atendê-lo agora. Tente mais tarde.* We are sorry, but we are unable to assist you at this time. Try again later.
A2	*Avariado. Pedimos desculpa pelo incómodo.* Out of service. We apologize for the inconvenience.

Table 2: Examples of short texts that only occur in the model exams of the A levels.

model exams consists of 32 texts. The distribution across levels differs from that of the texts of the private exams, with 20 of them belonging to the A levels. This is due to a type of reading comprehension exercise that includes several short texts and only occurs in the model exams of the A levels. Examples of these short texts are shown in Table 2.

3.2 Foundation Models

In terms of foundation models (Bommasani et al., 2021), we aim to extensively cover the models that are currently publicly available for Portuguese, independently of the language variety (Brazilian or European). They are described below.

3.2.1 BERTimbau

BERTimbau (Souza et al., 2020) is the most used Portuguese foundation model. It follows the original BERT architecture (Devlin et al., 2019), but it was trained on the Brazilian Web as a Corpus (brWaC) (Wagner Filho et al., 2018) solely for Masked Language Modeling (MLM). There are large and base variants of the model, with 335M and 110M parameters, respectively. There is also a distilled version of the model, obtained by applying the DistilBERT approach (Sanh et al., 2019) to the base variant.

3.2.2 BERTugues

BERTugues (Zago, 2023) improves on BERTimbau by being trained on a quality-filtered version of brWaC. Furthermore, it was also trained for Next Sentence Prediction (NSP). Additionally, its tokenizer includes emojis and discards characters that only very rarely occur in Portuguese. Contrarily to BERTimbau, BERTugues only has a base variant, with 110M parameters.

3.2.3 RoBERTa PT

RoBERTa PT (Santos et al., 2021) is a small version of RoBERTa (Liu et al., 2019) with 68M parameters trained on 10 million Portuguese sentences and 10 million English sentences from the OSCAR corpus (Suárez et al., 2019). It was trained by Santos et al. (2021) to be used in their study on automatic readability level assessment.

3.2.4 GPorTuguese-2

GPorTuguese-2 (Guillou, 2020) is a fine-tuned version of the English GPT-2 small model (Radford et al., 2019) on the Portuguese Wikipedia. It has 124M parameters. This was the model used as a foundation to achieve the highest performance in the study on automatic readability level assessment by Santos et al. (2021).

3.2.5 Albertina PT-*

Albertina PT-* (Rodrigues et al., 2023) is a family of models based on DeBERTa (He et al., 2021). There are models for both European Portuguese and Brazilian Portuguese. For each language variety, there are large and base variants of the model, with 900M and 100M parameters, respectively. The models for Brazilian Portuguese were trained on brWaC, while the ones for European Portuguese were trained on a combination of transcriptions of debates in the Portuguese Parliament, the Portuguese portions of European Parliament corpora, and the European Portuguese portion of the OSCAR corpus. Fine-tuned versions of these models currently achieve state-of-the-art performance on several NLP tasks in Portuguese.

3.3 Training & Evaluation Methodology

Starting with the evaluation metrics, we adopt accuracy, adjacent accuracy, and the macro F_1 score, which are some of the most common across previous studies on automatic readability level classification. Accuracy evaluates the precise identification of a text's readability level, while adjacent accuracy also considers neighboring levels, offering further insight into the identification of texts slightly easier or harder than the assigned level. Considering that the distribution of the texts across levels is not balanced, the macro F_1 score is also a relevant metric to understand whether the classifiers are biased toward the prediction of the majority classes.

The studies on automatic readability level assessment in European Portuguese described in Section 2 relied on cross-validation approaches to evaluation. As stated by Santos et al. (2021), cross-validation is not a common practice when training large neural models as it is a time-consuming process. Still, even though we defined a new test set for evaluation, we also relied on a 10-fold cross-validation approach to perform hyperparameter tuning and identify the top-performing foundation models for the task. This allows us to assess the performance of our models in an evaluation scenario that is similar to those of previous studies and to rely on the test set solely for assessing the generalization ability of the top-performing models.

In each fold of the cross-validation process, the foundation models are fine-tuned for 20 epochs. The weights of the best epoch are then selected according to the accuracy of the model. Considering that the cross-validation process generates 10 different fine-tuned models for each foundation model, we use them as an ensemble to generate the predictions for the test set. To aggregate the predictions of the multiple models, we experimented with approaches based on probability, ranking, and majority voting. We were not able to identify an approach that was clearly better than the others. Thus, we opted for averaging the class probabilities predicted by the multiple models.

To enhance robustness and mitigate the impact of randomness, we performed three independent experimental runs, each with a different random seed for the cross-validation splitting process. Then, we performed ten runs using the top-performing models to assess their generalization ability to the test set. Unless stated otherwise, the evaluation metrics are reported as both the average and standard deviation across these runs. All of the metrics are reported in percentage form.

3.4 Implementation Details

To train our models, we relied on the functionality offered by the HuggingFace's Transformers library (Wolf et al., 2020). We used the default values for most of the hyperparameters. However, we performed a grid search to identify appropriate values for the batch size and learning rate. For most foundation models, the best results were achieved using a batch size of 32 and a learning rate of 5×10^{-5}. One of the exceptions is GPorTuguese-2, which is highly influenced by padding. Thus, we used a batch size of 1. Furthermore, the best results were achieved using a lower learning rate of 1×10^{-5}. The other exception refers to the large versions of the Albertina PT-* models, which exhibited erratic behavior for larger values of the batch size and learning rate. Thus, we used a batch size of 16 and a learning rate of 1×10^{-5}.

4 Results

Considering that we use a cross-validation approach to identify the top-performing foundation models for automatic readability level classification in European Portuguese, in Section 4.1, we start by presenting and discussing the results achieved by the multiple foundation models in that scenario. Then, in Section 4.2, we take the best models and assess their generalization ability by analyzing their performance and errors on the test set.

4.1 Cross-Validation

Table 3 shows the results achieved by fine-tuning the multiple foundation models to the task. First of all, we can see that all models achieved an accuracy above 75%. In comparison, the best model in the study by Santos et al. (2021) achieved similar performance on the version of the dataset with 500 excerpts. This means that the additional training data we have available makes a significant impact on the performance of the models.

Looking into specific models, starting with BERTimbau, the most used foundation model for Portuguese, we can see that the performance of its three variants is as expected, with the large model performing better than the base one and the distilled version trading less than 1% performance for a reduced size and faster training and inference.

BERTugues was able to outperform the large version of BERTimbau despite having the same number of parameters as the base version. This was also observed by its author for other NLP tasks in Portuguese (Zago, 2023) and reveals the advantage of training foundation models on quality-filtered data and having a tokenizer that is more appropriate for the language.

RoBERTa PT, which is the smallest model used in our experiments, achieved performance similar

Model	Accuracy	Adjacent Accuracy	Macro F_1
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 Large	77.42±0.34	94.48±0.67	70.92±0.65
Albertina PT-BR Large	76.15±0.59	93.42±0.82	69.07±0.70
Albertina PT-PT Base	**81.77±0.44**	96.27±0.54	**76.17±1.01**
Albertina PT-BR Base	80.43±1.60	95.99±0.61	73.88±1.67

Table 3: Cross-validation results achieved by fine-tuning the foundation models to the task.

to that of the large version of BERTimbau in terms of accuracy and macro F_1 score and the highest adjacent accuracy overall. This can be justified by the improvements in the training process used by RoBERTa, such as dynamic masking (Liu et al., 2019). However, the pre-training on Portuguese sentences from the OSCAR corpus is also expected to have an impact, as the European variety of the language is considered as well.

GPorTuguese-2, the only foundation model of the GPT family used in our study, is one of the top-performing, ranking second in terms of every metric. Similarly to what was observed by Santos et al. (2021), it outperformed RoBERTa PT in terms of accuracy (by two percentage points in comparison to three in their study). The performance achieved using this model suggests that it is still a safe selection despite the existence of more recent foundation models. However, as its performance is impacted when dealing with padded inputs, it is not possible to take full advantage of modern hardware for its training, making it slower than fine-tuning the large variant of BERTimbau and nearly as slow as fine-tuning the large Albertina PT-* models, which have nearly nine times the number of parameters.

Looking into the results of the models in the Albertina PT-* family, we can see that the foundation models trained on data in European Portuguese outperform their Brazilian Portuguese counterparts. This confirms that the differences between the two varieties are relevant and impact how the difficulty level of a text is perceived.

Furthermore, among this family, we can find both the top and worst-performing models on this task. The large models that achieve state-of-the-art performance on several NLP tasks in Portuguese actually achieved the worst results in our experiments in terms of every metric. We argue that this is a case of overfitting, as these models are too large for the number of training examples available. Thus, we expect them to perform better given a sufficiently large and representative amount of training data. On the other hand, the base models are among the top performers on the task, achieving an accuracy above 80%.

Overall, the highest performance in the cross-validation scenario was achieved by fine-tuning the base version of the Albertina PT-PT model. The accuracy was 81.77% and the macro F_1 score was 76.17%. This also represents the lowest difference between both metrics across all models. On this subject, Santos et al. (2021) observed a difference of 13.60 percentage points when using RoBERTa PT and 6.72 percentage points when using GPorTuguese-2. Those values are reduced to 7.66 and 6.35 in our experiments, which suggests that the additional training data leads to less biased models. However, the difference between the metrics suggests that the models are still somewhat biased or that, at least, they have more difficulty in identifying examples of certain levels.

Table 4 shows the confusion matrices of the best runs of the two top-performing models. We can see that both models have a recall of at least 90% for the B1 level, which is both the middle level and the most prominent in the training dataset. On the other hand, the models seem to have some difficulties in distinguishing between the A levels. The main difference between the two models seems to be how they address the advanced levels. While GPorTuguese-2 seems to have some difficulties in distinguishing between the B2 and C1 levels, Al-

Albertina PT-PT Base						
		Predicted				
		A1	A2	B1	B2	C1
Actual	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

GPorTuguese-2						
		Predicted				
		A1	A2	B1	B2	C1
Actual	A1	73	17	2	0	0
	A2	29	127	1	0	0
	B1	3	7	218	6	6
	B2	0	0	6	28	15
	C1	0	0	3	15	42

Table 4: Confusion matrices of the best runs of the top-performing models in the cross-validation scenario: Albertina PT-PT Base (82.11% accuracy) and GPorTuguese-2 (81.60% accuracy).

bertina PT-PT Base seems to be more biased toward the prediction of the B1 level.

4.2 Generalization to the Test Set

Table 5 shows the performance of the two top-performing models in the cross-validation scenario when applied to the test set. We can see that the highest average performance is just 45.64% in terms of accuracy and 51.27% in terms of macro F_1 score, which reveals a lack of generalization ability by both models. Still, the GPorTuguese-2 model seems to generalize better than the base version of the Albertina PT-PT model in terms of accuracy and adjacent accuracy.

Among all the runs of the two models, we achieved a top performance of 50.00% in terms of accuracy, 84.38% in terms of adjacent accuracy, and 58.39% in terms of macro F_1 score. These results still represent a significant decrease in comparison to the performance achieved in the cross-validation scenario. Thus, it is important to assess the cause of this drop in performance when the models are applied to the test set.

Table 6 shows the confusion matrices of the best runs of Albertina PT-PT Base and GPorTuguese-2 when applied to the test set. We can see that the main difference observed between the two models in the cross-validation scenario can also be observed in this case. However, we can also see that both models predict several examples of the A levels as being of the B1 level. Without further information, one may be tempted to assume that the models are biased toward the prediction of the level that is predominant in the training data. However, by inspecting those examples, we found out that they correspond to the short texts, such as those shown in Table 2, that are exclusive to the model exams of the A levels. Their classification as B1 can be explained by the fact that, even though they are significantly longer, the shortest texts on the training data are of that level. Thus, the inability of the models to generalize their performance to this kind of text can be overcome by including more diverse kinds of text in the training data.

If those short texts are not considered, the average accuracy of the GPorTuguese-2 and Albertina PT-PT models improves to 76.84% and 72.63%, respectively. Although there is still a significant difference, these results are much closer to the performance in the cross-validation scenario. Due to space constraints and the size of texts, we are not able to show additional examples that are misclassified by the models. However, two examples are consistently misclassified. One of them is a dialog between two students about going to the library after class. It is of level A2 but is classified as level A1. The other is a description of the Erasmus+ program. It is of level C1 but is classified as being of one of the B levels. While the former can be explained by the simple vocabulary and the short sentences used in the dialog, the latter can be explained by the fact that the difficulty comes mainly from the length of the sentences. However, it is important to remember that the classification of texts by readability level is a task that is subjective and difficult even for humans (Branco et al., 2014a; Curto, 2014).

5 Conclusion

In this paper, we have addressed the automatic assessment of text readability level in European Portuguese. For that, we have explored the use of several foundation models and compared their performance when fine-tuned on textual data used for proficiency evaluation according to CEFR levels. Additionally, we have proposed a new publicly

Model	Accuracy	Adjacent Accuracy	Macro F_1
GPorTuguese-2	**45.63±3.02**	**81.56±0.99**	50.34±4.07
Albertina PT-PT Base	43.13±2.87	78.13±0.00	**51.27±3.93**

Table 5: Results achieved on the test set by the two top-performing models in the cross-validation scenario.

Albertina PT-PT Base						
		Predicted				
		A1	A2	B1	B2	C1
Actual	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

GPorTuguese-2						
		Predicted				
		A1	A2	B1	B2	C1
Actual	A1	3	0	5	0	0
	A2	1	3	8	0	0
	B1	0	0	5	0	0
	B2	0	0	0	2	1
	C1	0	0	0	1	3

Table 6: Confusion matrices of the best runs of the Albertina PT-PT Base (46.88% accuracy) and GPorTuguese-2 (50.00% accuracy) models on the test set.

available test set that promotes more comparable research on this subject.

Our experiments in a cross-validation scenario have shown that, considering the reduced amount of training data, the highest performance can be achieved by fine-tuning the base version of the recently released Albertina PT-PT model. However, for the same reason, the model has generalization issues when applied to kinds of text different from those that appear in its training data. Thus, similarly to many other NLP tasks in low-resourced languages, it is important to obtain more annotated data in order to train better models.

In future work, to mitigate the data scarcity problem, we intend to explore the use of data in the Brazilian variety of the language for training and assess whether the information provided by the additional data can outweigh the problems introduced by the differences between the two varieties. More broadly, we also want to explore the use of annotation data in other languages in combination with multilingual foundation models.

Additionally, considering the ordinal nature of the CEFR levels, we intend to assess whether there are benefits in addressing the problem as a regression task by fine-tuning the foundation models to output a continuous value instead of a specific level.

Still regarding potential approaches to the task, the emergence of large language models like ChatGPT (OpenAI, 2023) and LLaMa (Touvron et al., 2023), which exhibit commendable performance across various tasks, even in zero-shot scenarios, presents an enticing avenue to investigate.

Finally, considering the subjectivity of readability level assessment and its potential applications, it is important to make an effort towards the development of interpretable models for this task, in order to understand why a text is of a given level and how it can changed according to the proficiency level of the target audience.

Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

We would like to thank Camões, I.P. - Language Services Directorate for granting us access to the texts used in their exams and allowing us to use them to train our models.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S.

Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the Opportunities and Risks of Foundation Models. *Computing Research Repository*, arXiv:2108.07258.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014a. Assessing Automatic Text Classification for Interactive Language Learning. In *Proceedings of the International Conference on Information Society (i-Society)*, pages 70–78.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014b. Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology. In *Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, pages 256–261.

Miriam Cha, Youngjune Gwon, and H.T. Kung. 2017. Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 2003–2006.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, 54(5-6):340–359.

Pedro Curto. 2014. Classificador de Textos para o Ensino de Português como Segunda Língua. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2015. Automatic Text Difficulty Classifier. In *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, volume 1, pages 36–44.

Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, volume 1, pages 4171–4186.

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

William H. DuBay. 2004. *The Principles of Readability*. Impact Information.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic Text Difficulty Estimation Using Embeddings and Neural Networks. In *Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL)*, pages 335–348.

Luciana Forti, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci, and Stefania Spina. 2020. MALT-IT2: A New Resource to Measure Text Difficulty in Light of CEFR Levels for Italian L2 Learning. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 7204–7211.

Thomas François and Cédrick Fairon. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 466–477.

Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. AMesure: A Web Platform to Assist the Clear Writing of Administrative Texts. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP): System Demonstrations*, pages 1–7.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–407.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Maria José Grosso, António Soares, Fernanda de Sousa, and José Pascoal. 2011. QuaREPE: Quadro de Referência para o Ensino Português no Estrangeiro – Documento Orientador. Technical report, Direção-Geral da Educação (DGE).

Piere Guillou. 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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. Open Corpora and Toolkit for Assessing Text Readability in French. In *Proceedings of the Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 54–61.

Simon Jönsson, Evelina Rennes, Johan Falkenjack, and Arne Jönsson. 2018. A Component Based Approach to Measuring Text Complexity. In *Proceedings of the Swedish Language Technology Conference (SLTC)*, pages 58–61.

Nikolay Karpov, Julia Baranova, and Fedor Vitugin. 2014. Single-sentence Readability Prediction in Russian. In *Proceedings of the International Conference* *on Analysis of Images, Social Networks and Texts (AIST)*, pages 91–100.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Language Resources and Evaluation*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

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

Peter McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*, pages 3111–3119.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. In *Proceedings of the GermEval Workshop on Text Complexity Assessment of German Text*, pages 1–9.

Farah Nadeem and Mari Ostendorf. 2018. Estimating Linguistic Complexity for Science Texts. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

OpenAI. 2023. ChatGPT. https://chat.openai.com/.

Ildikó Pilán and Elena Volodina. 2018. Investigating the Importance of Linguistic Complexity Features Across Different Datasets Related to Language Learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. OpenAI Blog.

Robert Reynolds. 2016. Insights from Russian Second Language Readability Classification: Complexity-Dependent Training Requirements, and Feature Evaluation of Multiple Categories. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. *Computing Research Repository*, arXiv:2305.06721.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *Computing Research Repository*, arXiv:1910.01108.

Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. 2021. Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, pages 715–726.

Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. 2020. Automatic Classification of Text Complexity. *Applied Sciences*, 10(20):7285.

Carolina Evaristo Scarton and 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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Workshop on the Challenges in the Management of Large Corpora (CMLC)*, pages 9–16.

Yao Ting Sung, Wei Chun Lin, Scott Benjamin Dyson, Kuo En Chang, and Yu Chia Chen. 2015. Leveling L2 Texts through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Hugo Touvron, 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, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Computing Research Repository*, arXiv:2302.13971.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: a New Open Resource for Brazilian Portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4339–4344.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment Toolkit. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1217–1233.

Thomas Wolf, 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, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Kevin Yancey, Alice Pintard, and 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.

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