Neural Text Categorization with Transformers for Learning Portuguese as a Second Language

Rodrigo Santos[1](✉), João Rodrigues[1], António Branco[1], and Rui Vaz[2]

[1] NLX—Natural Language and Speech Group, Department of Informatics, Faculdade de Ciências, University of Lisbon, 1749-016 Campo Grande, Lisbon, Portugal
{rsdsantos,jarodrigues,antonio.branco}@fc.ul.pt
[2] Camões I.P. Instituto da Cooperação e da Língua,
Av. da Liberdade 270, 1250-149 Lisbon, Portugal
rvaz@camoes.mne.pt

Abstract. We report on the application of a neural network based approach to the problem of automatically categorizing texts according to their proficiency levels and suitability for learners of Portuguese as a second language. We resort to a particular deep learning architecture, namely Transformers, as we fine-tune GPT-2 and RoBERTa on data sets labeled with respect to the standard CEFR proficiency levels, that were provided by Camões IC, the Portuguese official language institute. Despite the reduced size of the data sets available, we found that the resulting models overperform previous carefully crafted feature based counterparts in most evaluation scenarios, thus offering a new state-of-the-art for this task in what concerns the Portuguese language.

Keywords: Readability classification · Language proficiency · Neural networks · Deep learning · Portuguese

1 Introduction

Learning and teaching an idiom as a second language is a challenge for students and teachers. While the former struggle with the acquisition of a new language, the latter have, among other things, to gather and create study materials that efficiently support the acquisition of that new cognitive skill.

Automatic Text Difficulty Classification, also known as Readability Assessment, can ease the work for both students and teachers as it helps to determine the level of difficulty of a text, since learning from a text that is too easy renders few benefits, and learning from a text that is to hard may frustrate the students.

Despite the usefulness of these systems for any language, most research and data is in and for English—as it is common in the field of Natural Language Processing (NLP)—leaving most languages unsupported.

With the scarcity of data for the vast majority of languages, the development of language tools for the extraction of features used on the creation of a classifiers for text difficulty level becomes an even greater challenge.

G. Marreiros et al. (Eds.): EPIA 2021, LNAI 12981, pp. 715–726, 2021.
https://doi.org/10.1007/978-3-030-86230-5_56

The work presented in this paper focuses on the classification of proficiency levels of Portuguese texts, a language with considerably few resources labelled according to the CEFR (Common European Framework of Reference for Languages) [19] levels, as this is the categorization used by the Portuguese official language agency Camões IP[1] in its teaching and certification activities. The CEFR levels classify text difficulty into six reference levels of increasing proficiency and difficulty, viz. A1, A2, B1, B2, C1, C2, and is widely accepted as the European standard for grading language proficiency.

Like in almost any other NLP task, automatic text classification has recently seen a boost in performance with the introduction of the neural network, deep learning architecture known as Transformer [44], more precisely through the use of gigantic deep language models that make use of this architecture and are pre-trained on very large data sets of raw text.

This unsupervised pre-training step helps the model to learn an inner representation of the language and is used to complement and alleviate the very short volume of labelled data sets that are specific for any given task, and are used for the subsequent fine-tuning of the model.

The research question of the present paper aims to answer is: while using Portuguese as an example and case study of a low-resourced language with respect to labelled data sets for language proficiency, can the new transformer based language models be shown to support a solution that over-performs the state of the art provided by feature based models for proficiency level classification on low-resource scenarios?

A positive answer to this question can benefit the low resourced languages, which are the vast majority, with few resources and that are not technologically prepared to have the tools that are required for the extraction of features for text difficulty classification.

We find that the GPT-2 [29] and RoBERTa [25] language models support very competitive scores in comparison to a feature based classifier, with GPT-2 surpassing the feature based approach and setting a new state-of-the-art performance for the Portuguese language.

Another major outcome of the research reported in this paper is the resulting classifier, which we make available as the online tool LX-Proficiency to support students of Portuguese as well as anyone interested.[2]

The remainder of this document is organized as follows: Sect. 2 presents the relevant previous work; Sect. 3 introduces the corpora used throughout our experiments; Sect. 4 summarizes the Transformer architecture and the language models used; Sect. 5 describes the implementation of the feature-based model as well as the deep language models; Sect. 6 presents and discusses the results obtained. Finally, Sect. 7 closes this document with concluding remarks.

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

[2] https://portulanclarin.net/workbench/lx-proficiency.

2 Related Work

Readability assessment has a long tradition of research on a wide range of different readability indexes (cf. overviews in [13,18]). Recently, text difficulty classification or readability assessment has been addressed in tasks such as automatic proficiency level classification. This task aims to classify excerpts of texts in accordance to a given range of proficiency levels (typically as set up in CEFR). In addition to unsupervised readability indexes, most of the recent work has resorted to a wide range of features to train machine learning classifiers: count-based, lexical, morphological, syntactic, and semantic [24,28].

In spite of the scarceness of data, within the paradigm of readability indexes and machine learning classifiers, a few languages have been addressed: Chinese [41], Dutch [42], Estonian [43], French [22], German [23], Italian [21,37], Russian [30] and Swedish [34], among others.

The authors of [26] study the impact of neural networks in this task and achieve some success using BERT [17], HAN [47], and a Bi-LSTM [38] with accuracy scores ranging from 78.72% to 85.73% on three English datasets. Despite this success, they obtain good performance only for the English language, and the same type of models underperform when training on Slovenian with a smaller data set (52.77% accuracy).

Readability assessment in the Portuguese language is a research domain still largely untapped. However, it is worth mentioning the closely related work reported in [1], on a wide experimental space for Portuguese readability assessment for text simplification; in [31], on the task of automatic scoring texts produced by learners of Portuguese as a second language; and in [36], on measuring the impact of readability features to detect fake news.

Regarding the published research specifically on text difficulty assessment in Portuguese, it was reported in three papers by two teams, namely [8,9,16], which classify texts into CEFR levels and use corpora from Camões IP.

The first paper [9] makes use of a corpus with 114 labelled excerpts and four unsupervised metrics in order to classify the texts. The metrics are: the Flesch Reading Ease index [20] (27.03% accuracy); the lexical category density in the proportion of nouns (22.97% accuracy); the average word length in the number of syllables per word (29.73% accuracy); and the average sentence length in the number of words per sentence (19.09% accuracy). Their purpose was the creation of a tool to help language learners and teachers of Portuguese to assess the level of a text. Accordingly, they don't merge the features that were extracted—which could help obtain a higher performance score—as it could blur the interpretability of the tool by its users.

The second paper [8] includes a re-evaluation by human experts of the tool presented in the first paper.

Finally, the third paper [16] makes uses of a second corpus that was double in size, with 237 labelled excerpts (including the 114 excerpts used in the first two papers), and focused on extracting 52 features from the text to experiment with various machine learning models, which deliver the best performance by resorting to LogitBoost (75.11% accuracy).

Table 1. Example of the Portuguese corpus

A Célia tem 15 anos e é verdadeiramente uma pessoa da era digital. Gosta muito de informática e de novas tecnologias, mas também de viagens	A1
O Ballet Clássico alia o movimento dançado ao sentido de musicalidade, inspirando-se no universo das danças populares e palacianas	B2

3 Corpus

In its certification activities, Camões IP is responsible for running language exams on Portuguese as a second language worldwide, and thus assessing the correct difficulty level of the text excerpts for each exam is crucial.

The corpus used in the present paper was provided by Camões IP. It has 500 excerpts of Portuguese news, books and articles (including the data sets used by the three previous papers mentioned above), which are labelled with one of five CEFR levels, namely A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), and C (advanced and proficient).

Table 1 showcases sentences from two excerpts in this corpus, separated by three levels of difficulty. We can observe that to correctly classify them, one as to take into account various factors such as sentence length or the type of vocabulary.

While this 500 texts corpus is pretty small for deep learning standards, it represents a great improvement over data sets available for previous work, doubling the 237 texts previously available to [16] and almost quintupling the 114 texts available to [8,9].

Table 2 contains the global statistics for the corpora used in this work, and Table 3 discriminates the proportion of excerpts in each class.

In order to allow for comparison to previous work, the data used was divided into 5 subsets: (i) a set that encompasses all the 500 texts that are available, termed c500 for ease of reference; (ii) a balanced set where every class has 45 texts randomly selected inside each class—capped by the size of the smallest class, B2, in c500—making a total of 225 texts, termed c225bal for ease of reference; (iii) a set that approximates the corpus of [8,9] with 114 texts, termed c114; (iv) a set with 88 texts from [8,9] consisting of a re-annotated version of subset of c114 with some texts removed due to insufficient agreement between annotators, termed c88r; and finally (v) a set that approximates the corpus used in [16] with 237 texts, which contains c114, termed c237 for ease of reference;

In Table 2, we can see that the number of sentences in each corpus closely follows the trend of number of excerpt/texts with the exception being c225bal with more sentences than c237 that has more texts.

The c225bal corpus is also the corpus with highest average of tokens and sentences per excerpt showcasing that each excerpts are in average here larger than in any other corpus.

Table 2. Corpora statistics

Corpus	Excerpts	Tokens	Av. tokens/excerpts	Sentences	Av. sentences/excerpts
c500	500	89,749	179.50	5,647	11.29
c225bal	225	49,734	221.04	2,999	13.33
c237	237	37,592	158.62	2,122	8.95
c114	114	12,875	112.94	677	5.94
c88r	88	10,793	122.65	588	6.68

Table 3. Class distribution

Corpus	A1		A2		B1		B2		C	
	Num.	Per.	Num.	Per.	Num.	Per.	Num.	Per.	Num.	Per.
c500	80	16%	135	27%	184	36.8%	45	9%	56	11.2%
c225bal	45	20%	45	20%	45	20%	45	20%	45	20%
c237	29	12.2%	39	16.5%	136	57.4%	14	5.9%	19	8%
c114	11	9.6%	11	9.6%	72	63.2%	8	7%	12	10.5%
c88r	30	34.1%	17	19.3%	23	26%	11	12.5%	7	8%

Table 3 help also to show how imbalanced c114 and c237 are, with the class with the highest percentage B1 having 63.2% and 57.4% of all texts, respectively, and the class with lowest percentage having as little as 7% and 5.9%, respectively

The last line of Table 3 presents the distribution of c88r. Given this corpus is a re-annotated version of a subset of c114, we can see that many of the texts previously classified as B1 in c114 have here a different label, with every other class growing in size with the exception of C. This can negatively affect a model trained with c114 that despite being able to achieve a good performance score, in reality the model is adjusted to data that is wrongly labeled and has a big bias towards one class. This same issue might be prevalent in c237, and it it only somewhat mitigated in c225bal, due to the balance of every class, and in c500, as its size may help mitigate the wrong classification.

4 Transformer Models

The introduction of the deep learning architecture Transformer [44] has produced a revolution in the field of Natural Language Processing (NLP). Where previously sets of rules, features, and various machine learning models were used, they have been successfully replaced with a variant of the Transformer architecture.

These Transformer variants are machine learning algorithms that obtain state-of-art performance on a wide range of Natural Language Understanding and Generation tasks. They are neural network algorithms that encapsulate a tokenizer, contextual embeddings and a task-specific prediction algorithm.

Typically the deep learning pipeline consists of string tokenization, converting raw text to a sparse index encoding, followed by a transformation to sparse indices resorting to several neural network layers (representing a contextual embedding), and finally, a head layer outputs to a task-specific prediction (e.g. language modeling, sequence classification, question answering or conditional generation among others).

In terms of architecture, the Transformer [44] extrapolates the idea underlying the Attention Mechanism [2] and creates a sequence-to-sequence encoder-decoder model that relies almost only on Attention layers, creating a model that is both better and faster at dealing with language processing tasks than the previous Recurrent Deep Networks.

In this work we use two Transformer based models, namely the GPT-2 [29] model and the RoBERTa [25] model. The tokenizers from both models rely on a statistical analysis from the training corpus using subwords units. More specifically, they use a byte-level Byte-Pair Encoding (BPE) vocabulary [39].

4.1 GPT-2

The GPT-2 is an autoregressive neural network that makes use of the decoder side of the Transformer model and its training objective is to decode the next token (word or piece of word) in a sentence.

More precisely, given an input sequence $x_{1:n-1}$, it learns by predicting the next word:

$$x_{1:n-1} \Rightarrow x_n \tag{1}$$

Internally the model uses a mask-mechanism to make sure the predictions for the target token n only use the inputs from x_1 to x_{n-1} but not the future tokens, which means that for the prediction of each token the model only has access to the tokens on the left of the target token. This means that the model is pre-trained on raw texts only, with no need for human-labeled data. While the model thrives on generation tasks, it can be used to extract features that can be used on various downstream tasks.

During fine-tuning, a classification head is added to the top of the model. The head performs a sequence classification for each input sequence $x_{1:N}$ and gives a possible output y from a class set C:

$$x_{1:N} \xrightarrow{\text{outputs}} y \in C \tag{2}$$

We implemented the GPT-2 classification model resorting to the open-source library *Transformers* [46], with a 12 layers and 12 attention-heads model architecture, totaling 124M parameters, and initializing the model with a model fine-tuned from English to Portuguese.[3]

4.2 RoBERTa

Like GPT-2, the RoBERTa model makes use of part of the Transformer model, only this time it is the encoder side on the Transformer that is used. The

[3] https://huggingface.co/pierreguillou/gpt2-small-portuguese.

RoBERTa model is an improvement upon the BERT model [17] as it does not use the NSP (Next Sentence Prediction) training objective, and uses more data, longer sequences, and a bigger batch size than the original BERT model.

RoBERTa has the train objective named MLM (Masked Language Model), where a word at random is masked in the sentence and the model is asked to predict what was the word that was masked. In particular, this model receives an input token sequence $x_{1:N}$ and trains the model by predicting a masked word x_n that was swapped at random by a mask token (e.g., <MASK>):

$$x_{1:N\setminus n} \Rightarrow x_n \tag{3}$$

Internally, the model has access to every word on the left and on the right of the masked word, creating a stronger context than in the GPT-2 model (only words to the left) to predict the word. Just like GPT-2 during fine-tuning, a classification head is added to the top of the model (see Eq. 2).

Since hugging face has no pre-trained RoBERTa model in Portuguese, we trained a new RoBERTa model with 6 layers and 12 attention-heads, totaling 68M parameters, on 10 million Portuguese sentences and 10 million English sentences from the Oscar corpus.[4]

5 Implementation

In order to compare with previous work, we re-implemented the classifier from [16], having gathered information from this paper, which is based on the dissertation [15], from one of the authors, and from the tool's website[5].

We trained the classifier using 10-cross fold validation, used the same features, and the same classifier LogitBoost.[6] The paper does not indicate the parameters used, so we used the default parameters. With the average of 3 runs, we found a performance score (74.12%) that is in line with the score (75.11%) reported in the reproduced paper [16], both presented in the Table 4. The difference of 0.99% in accuracy makes us confident that both the reproduced classifier and reproduced corpus are close to the ones in the original paper.

As mentioned above, in this work we make use of the GPT-2 and RoBERTa models for classifying Portuguese text into one of the five CEFR proficiency levels. We fine-tuned both models on the five corpora, with both models using a batch size of 1; 5 epochs for the c88r and c225bal, 10 epochs for c114, c237 and c500; and using a learning rate of $2e-5$ for GPT-2 and $1e-5$ for RoBERTa.

We trained/fine-tuned each model using 10-fold cross validation. While this method is not usual for neural networks, mainly because it is very time consuming, we used it in order to allow for comparison with the scores obtained in previous work. Every model is trained three times and the performance scores reported are an average of these three runs.

[4] https://oscar-corpus.com/.

[5] https://string.hlt.inesc-id.pt/demo/classification.pl.

[6] https://logitboost.readthedocs.io/.

Table 4. Performance (accuracy)

Model	c114	c88r	c237	c500	c225bal
Unsupervised indexes [9]	21.82%	29.73%	-	-	-
Feature-based LogitBoost [16]	-	-	75.11%	-	-
Our LogitBoost reproduction	**86,84%**	48,86%	74,12%	68,60%	59,70%
GPT-2	84,21%	55,68%	**76,23%**	**75,62%**	**65,48%**
RoBERTa	85,32%	**57,83%**	75,45%	72,50%	63,19%

6 Evaluation and Discussion

While accuracy allows for a quick and intuitive grasp of the performance, in scenarios like ours where classes are severely unbalanced, other metrics can be more sensible. Hence, we complement accuracy with macro-averaged f1 score and quadratic weighted kappa as in both these metrics all classes contribute equally regardless of how often they appear in the test set. Table 5 presents the performance scores for LogitBoost, GPT-2 and RoBERTa.

The absolute highest performance score (86.84%) is obtained for the c114 corpus using LogitBoost. While one might consider that this model trained with this corpus is the best performing model, a case can be made that in reality the high imbalance of the corpus, mainly in the B1 class, creates a heavily biased model that will performs poorly in a real world scenario. The same argument can be provided for c237. Like c114, this data set is also heavily imbalanced.

The most simplistic baseline model that always answered with the majority class (B1) would already achieve as much as 57.4% and 63.2% accuracy with c237 ans c114 respectively (cf. Table 3). Despite this, the c237 corpus is our comparison gateway to the work of [16] (scoring 75.11%), and both the GPT-2 (with 76.23%) and RoBERTa (75.45%) models achieve higher accuracy than it, with the GPT-2 model even beating the other models for c237 in all the three evaluation metrics (Table 5).

The corpus supporting the worst performance scores is c88r, which is not surprising due to its reduced size. Here no model has a clear advantage over the others, with each model being better than the other two in one of the three metrics. Both c88r and c114 are the gateway to comparison with [8,9], which are outperformed by the other models, mainly due to the unsupervised indexes and the rudimentary algorithm used, viz. linear regression.

Finally, the novel data sets presented in this work c500 and c225bal support their best performance with GPT-2, and both appear as strong candidates for a real world application.

Given its the largest balanced corpus, with all classes with equal size, the latter presents itself as a fairer model for all classes. Accordingly, it can be seen as providing the best sensible scores to compare the performance of the various models—with GPT-2 outperforming the other models—and thus the reference scores for this task for the Portuguese language given the labelled data available

Table 5. Performance (accuracy, macro-f1, quadratic weight kappa)

Model	c114			c88r			c237			c500			c225bal		
	acc	f1	qwk	acc	f1	qwk	acc	f1	qwk	acc	f1	qwk	acc	f1	qwk
LogitBoost	**86,84**	**0,737**	**0,898**	48,86	**0,429**	0,720	74,12	0,553	0,735	68,60	0,643	0,791	59,70	0,595	0,809
GPT-2	84,21	0,675	0,793	55,68	0,348	**0,736**	**76,23**	**0,556**	**0,760**	**75,62**	**0,689**	**0,859**	**65,48**	**0,649**	**0,879**
RoBERTa	85,32	0,615	0,792	**57,83**	0,322	0,691	75,45	0,510	0,709	72,50	0,589	0,826	63,19	0,562	0,848

at present. In turn, due to its larger size, the former has also an interesting advantage as it has seen and learned from a wider range of examples and may more closely represent the distribution of data in the real world. Here again, GPT-2 is by far the best performing model, in all evaluation metrics.

7 Conclusion

The results reported in this paper show that, despite the very small dimension of the labeled data available and the known need of neural methods for large data sets, the neural-based transformer based language models are capable of performing on par with non neural models trained on features in the task of text difficulty classification, even achieving state-of-the-art performance for the Portuguese language. These results thus demonstrate that good performance on the task can be achieved if one has a small corpus, and can thus dispense with auxiliary language tools for feature extraction needed by non neural learning approaches previously used. This comes as good news for under-resourced languages that do not have the language resources and tools needed for the extraction of the relevant features.

Moreover, we offer access to the GPT-2 model trained with the c500 corpus, as we deem it as the model that more likely follows closely the distribution of classes in the real usage scenario of the students of Portuguese as second language applying to certification from Camões IP. It underlies the online service LX-Proficiency,[7] which can be freely accessed online from the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language[8].

For future work we would like to study the impact of knowledge transfer from other languages that have more resources (e.g. English), as well as methods to synthetically increase the size of the training set.

Acknowledgements. The work leading to the research results reported in this paper were mostly supported by Camões I.P. Instituto da Cooperação e da Língua. It was also partially supported by PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

[7] https://portulanclarin.net/workbench/lx-proficiency.

[8] The PORTULAN CLARIN workbench comprises a number of tools that are based on a large body of research work contributed by different authors and teams, which continues to grow and is acknowledged here: [3–7,10–12,14,27,32,33,35,40,45].

References

1. Aluisio, S., Specia, L., Gasperin, C., Scarton, C.: Readability assessment for text simplification. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–9 (2010)
2. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
3. Barreto, F. et al.: Open resources and tools for the shallow processing of Portuguese: the TagShare project. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 1438–1443 (2006)
4. Branco, A., Henriques, T.: Aspects of verbal inflection and lemmatization: generalizations and algorithms. In: Proceedings of XVIII Annual Meeting of the Portuguese Association of Linguistics (APL), pp. 201–210 (2003)
5. Branco, A., Castro, S., Silva, J., Costa, F.: CINTIL DepBank handbook: Design options for the representation of grammatical dependencies. Technical report, University of Lisbon (2011)
6. Branco, A., et al.: Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), pp. 1810–1815 (2010)
7. Branco, A., Nunes, F.: Verb analysis in a highly inflective language with an MFF algorithm. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS (LNAI), vol. 7243, pp. 1–11. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28885-2_1
8. Branco, A., Rodrigues, J., Costa, F., Silva, J., Vaz, R.: Assessing automatic text classification for interactive language learning. In: International Conference on Information Society (i-Society 2014), pp. 70–78 (2014)
9. Branco, A., Rodrigues, J., Costa, F., Silva, J., Vaz, R.: Rolling out text categorization for language learning assessment supported by language technology. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS (LNAI), vol. 8775, pp. 256–261. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09761-9_29
10. Branco, A., Rodrigues, J., Silva, J., Costa, F., Vaz, R.: Assessing automatic text classification for interactive language learning. In: Proceedings of the IEEE International Conference on Information Society (iSociety), pp. 72–80 (2014)
11. Branco, A., Silva, J.: A suite of shallow processing tools for Portuguese: LX-suite. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 179–182 (2006)
12. Costa, F., Branco, A.: Aspectual type and temporal relation classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 266–275 (2012)
13. Crossley, S.A., Skalicky, S., Dascalu, M., McNamara, D.S., Kyle, K.: Predicting text comprehension, processing, and familiarity in adult readers: new approaches to readability formulas. Discourse Process. **54**, 340–359 (2017)
14. Cruz, A.F., Rocha, G., Cardoso, H.L.: Exploring Spanish corpora for Portuguese coreference resolution. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 290–295 (2018)
15. Curto, P.: Classificador de textos para o ensino de português como segunda lıngua. Master's thesis, Instituto Superior Técnico-Universidade de Lisboa, Lisboa (2014)

16. Curto, P., Mamede, N., Baptista, J.: Automatic text difficulty classifier. In: Proceedings of the 7th International Conference on Computer Supported Education, vol. 1, pp. 36–44 (2015)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
18. DuBay, W.H.: The Principles of Readability. Impact Information, Costa Mesa (2004)
19. Council for Europe, Council for Cultural Co-operation, E.C., Division, M.L.: Common European Framework of Reference for Languages: learning, teaching, assessment (2001)
20. Flesch, R.: How to Write Plain English: A Book for Lawyers and Consumers. Harpercollins, New York (1979)
21. Forti, L., Grego G., Santarelli, F., Santucci, V., Spina, S.: MALT-IT2: a new resource to measure text difficulty in light of CEFR levels for Italian l2 learning. In: 12th Language Resources and Evaluation Conference, pp. 7206–7213 (2020)
22. François, T., Fairon, C.: An "AI readability" formula for French as a foreign language. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 466–477 (2012)
23. Hancke, J., Meurers, D.: Exploring CEFR classification for German based on rich linguistic modeling. In: Learner Corpus Research, pp. 54–56 (2013)
24. Jönsson, S., Rennes, E., Falkenjack, J., Jönsson, A.: A component based approach to measuring text complexity. In: The Seventh Swedish Language Technology Conference (SLTC-18), Stockholm, Sweden, 7–9 November 2018 (2018)
25. Liu, Y., et al.: Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
26. Martinc, M., Pollak, S., Robnik-Šikonja, M.: Supervised and unsupervised neural approaches to text readability (to be published)
27. Miranda, N., Raminhos, R., Seabra, P., Sequeira, J., Gonçalves, T., Quaresma, P.: Named entity recognition using machine learning techniques. In: EPIA-11, 15th Portuguese Conference on Artificial Intelligence, pp. 818–831 (2011)
28. Pilán, I., Volodina, E.: 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, pp. 49–58 (2018)
29. Radford, A., et al.: Better language models and their implications. OpenAI Blog (2019). https://openai.com/blog/better-language-models
30. Reynolds, R.: Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 289–300 (2016)
31. del Río, I.: Automatic proficiency classification in l2 Portuguese. Procesamiento del Lenguaje Nat. **63**, 67–74 (2019)
32. Rodrigues, J., Costa, F., Silva, J., Branco, A.: Automatic syllabification of Portuguese. Revista da Associação Portuguesa de Linguística (1), 715–720 (2020)
33. Rodrigues, J., Branco, A., Neale, S., Silva, J.: LX-DSemVectors: distributional semantics models for Portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 259–270. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_27

34. Santini, M., Jönsson, A., Rennes, E.: Visualizing facets of text complexity across registers. In: Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), pp. 49–56 (2020)
35. Santos, R., Silva, J., Branco, A., Xiong, D.: The direct path may not be the best: Portuguese-Chinese neural machine translation. In: Proceedings of the 19th EPIA Conference on Artificial Intelligence, pp. 757–768 (2019)
36. Santos, R., et al.: Measuring the impact of readability features in fake news detection. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 1404–1413 (2020)
37. Santucci, V., Santarelli, F., Forti, L., Spina, S.: Automatic classification of text complexity. Appl. Sci. **10**, 7285 (2020)
38. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
39. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725 (2016)
40. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-box robust parsing of Portuguese. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), pp. 75–85 (2009)
41. Sung, Y.T., Lin, W.C., Dyson, S.B., Chang, K.E., Chen, Y.C.: Leveling l2 texts through readability: combining multilevel linguistic features with the CEFR. Mod. Lang. J. **99**, 371–391 (2015)
42. Tack, A., François, T., Desmet, P., Fairon, C.: NT2Lex: a CEFR-graded lexical resource for Dutch as a foreign language linked to open Dutch wordnet. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 137–146 (2018)
43. Vajjala, S., Loo, K.: Automatic CEFR level prediction for Estonian learner text. In: Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning, pp. 113–127 (2014)
44. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010 (2017)
45. Veiga, A., Candeias, S., Perdigão, F.: Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (2011)
46. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
47. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)