Performance of Transformer Models in Readability Assessment

David Sachelarie[1]

Supervisor: Maria Soledad Pera[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 24, 2023

Name of the student: David Sachelarie
Final project course: CSE3000 Research Project
Thesis committee: Maria Soledad Pera, Pradeep Murukannaiah

Abstract

Transformer models have proven to be effective tools when used for determining the readability of texts. Models based on pre-trained architectures such as BERT, RoBERTa, and BART, as well as ReadNet, a transformer model which is dedicated to readability assessment, have shown some very promising results. However, there is a lack of research focused on comprehensively analyzing these models' performance at a more granular level. Moreover, GPT-2, a member of the very popular GPT transformer family, has never been adapted to and tested in readability assessment. The work presented in this paper fills these knowledge gaps by analyzing the behavior of the five aforementioned models and reflecting on their performance on separate classes of text difficulty. Seeing how they perform on texts of various complexity levels is vital to understanding their behavior and limitations, which will in turn further the knowledge of the situations in which each readability tool achieves the most optimal results.

1 Introduction

Understanding how well can a text be understood by its readers, or, in other words, its **readability** level, is of major importance for many conveyors of information. Study materials, newspaper articles, and governmental publications, to name but a few, greatly depend on controlling readability [5]. Traditionally, a text's readability level has been calculated through the use of various formulas [1; 13], but, with improvements in machine learning and natural language processing, it was only a matter of time until several novel approaches were proven to be more effective means of performing readability assessment [12]. Among the aforementioned novel techniques are the **transformer models**, neural networks which rely "entirely" on attention mechanisms, "dispensing with recurrence and convolutions" [18]. Several of these models have been found to be performing better than simple machine learning approaches when evaluating text difficulty [9; 13].

Transitioning from well-known, relatively simple formulas to machine learning models, and specifically transformer models, is a challenging process in readability assessment [3]. That is because the latter tend to be seen as opaque black-box solutions that behave in ways which are hard to comprehend or explain [8]. This point of view is perfectly justified, since, even though models based on the **BERT** [6], **RoBERTa** [11], and **BART** [10] transformer architectures have proven to be performing well in readability assessment [9], it is still unclear which classes of text difficulty each model performs best and worst on. Researching the limitations and behavioral characteristics of neural natural language processing models used for assessing a text's readability is thus essential to accelerating these new models' widespread introduction in this field.

Another transformer model which showcased promising results in previous research is **ReadNet** [13], a model which is not based on any pre-trained architecture, being designed exclusively for readability assessment. To our knowledge, there are no empirical works which compare and contrast ReadNet's performance to the one of the three models described above, a situation which this paper aims to remedy. Even though all of these models' performance is very similar on paper [9; 13], running them in the same environment (and, most importantly, on the same dataset split) is important for revealing subtle differences in their performance and ultimately deciding which one is better.

Considering the rising popularity of ChatGPT and the sheer number of GPT-related models which have emerged in the past few years [2], we couldn't overlook the fact that this family of transformers has never been tested in readability assessment. To remedy this situation, we want to introduce a readability model which is built upon the easily fine-tunable **GPT-2** transformer architecture [16]. Since readability assessment is part of the broader field of natural language processing, and GPT-related models have proved to be very popular and effective tools when used for tackling a wide range of natural language processing tasks [2; 15], we surmised that there is potential for GPT-2 to be performing well in readability assessment.

To advance research pertaining the better understanding of how transformer architectures fare when applied to the task of readability estimation, conducting a comprehensive analysis of the aforementioned models' behavior on a very granular dataset is essential. Such research which, to the best of our knowledge, has never been carried out before, will make up the core of the work presented in this paper.

Our work aims to answer the following main research question: *What are the strengths and weaknesses of several transformer models used for readability assessment?*. For answering this question, the following sub-questions have to be considered:

- How do the models perform on every class of text difficulty the dataset provides?
- How significant are the models' performance differences between text difficulty classes?
- How do the results compare to the baseline?

The research question is addressed through an empirical exploration of the performance of five models: ReadNet, and fine-tuned BERT, RoBERTa, BART, and GPT-2 models. They are trained and evaluated on a granular corpus, and their predictions are considered in the context of the samples' difficulty. The strengths and weaknesses are then inferred from each model's performance differences on the labels.

The work presented in this paper contributes to the field of readability assessment by furthering the understanding of how well transformer models perform in different situations. Having this knowledge is essential for deciding when to use each readability tool in order to get the most optimal results. Since our hope is that future research will expand upon our work, thus making this understanding even more comprehensive, we made the bulk of our code freely available, so that our research can be reproduced as easily as possible.

The rest of this paper is organized as follows: in Section 2, we document the research done by past papers, in Section 3

we present the research methodology, Section 4 covers the results which were reached, Section 5 provides a reflection on the ethical aspects of the research, section 6 contains a discussion on the experiment's outcomes, while the limitations, ideas for future work, implications, and conclusions are presented in Section 7.

2 Related Work

The research we conducted expands upon the work presented by two papers.

The first paper [9] introduces the "hybrid models", which are transformer models (BERT, RoBERTa, BART, and XL-Net) combined with traditional machine learning techniques such as Random Forest and Gradient Boosting. This paper is not relevant for our research because of the hybrid models though, but since it discusses the performance of BERT, RoBERTa, and BART on three corpora, one of them being WeeBit. This paper offers a useful, though limited overview on how well models based on pre-trained architectures perform in readability assessment.

The second paper analyzes the performance of a standalone transformer model named ReadNet [13]. The model is evaluated on WeeBit and two other corpora, and is compared to several machine learning algorithms. What is lacking from this research though is a head-to-head encounter between ReadNet and other transformer models, being thus unclear whether this model could compete against its powerful pre-trained peers.

3 Experimental Setup

In this section, we present the models, data, and metrics considered in our study, along with the exploration protocol we followed.

3.1 Models

We evaluate the performance of readability assessment models adapted from four pre-trained architectures:

BERT - A "multi-layer bidirectional" pre-trained transformer. Its original objectives are masked language modeling and next sentence prediction, but BERT can easily be fine-tuned to support a variety of other tasks [6].

RoBERTa - A BERT model with an "improved training procedure" [11].

BART - A pre-trained transformer model which "maps a corrupted document to the original document it was derived from". Just like BERT and RoBERTa, it can also be fine-tuned for supporting other tasks [10].

GPT-2 - An improved version of the original GPT model [16].

In order to use the four architectures listed above for readability assessment, we added a text classification head on top of the pre-trained models. The resulting fine-tuned models were compared against each other but also against **ReadNet**, a transformer model dedicated to readability assessment. Even though this transformer is not a large language model (LLM) like the architectures presented above, the fact that it leverages "specific features indicating the readability" of text should make it effective in readability assessment [13].

The **Flesch-Kincaid grade level**, one of the most popular formulas used for readability assessment, was used as a baseline for comparing the performance of the aforementioned transformer models. It predicts the amount of grade levels a person needs to follow to understand a given text [7]. For example, someone with a fourth-grade reading ability should be able to comprehend a text with a Flesch-Kincaid grade level score of 4.

3.2 Data

WeeBit is one of the most granular, largest, and most widely used datasets for readability assessment [9; 13]. It is targeted towards children and teenagers and classifies its data based on their age, in five categories: 7-8, 8-9, 9-10, 11-14 and 15-16 years old [17]. In order to perform unbiased training, we downsampled WeeBit to 2235 samples, 447 per difficulty level, as can be seen in Table 1. This is due to two reasons.

Firstly, some of the models didn't support certain special characters, so the files containing them had to be removed from the dataset. We named the trimmed corpus, whose composition is presented in Table 1, WeeBit-NoSpecialChar.

Secondly, irrelevant information "strongly correlated with the target labels" had to be removed from WeeBit-NoSpecialChar's samples [12]. We did not find any specific instructions on how to do that in previous research, so we proceeded to parse samples manually in order to discover some patterns. It turned out that paragraphs with useful text never had any "garbage" mixed in, the unwanted text snippets appearing in separate paragraphs. Moreover, with only one exception which was treated separately, all of these unwanted paragraphs consisted of less than 2 complete declarative sentences. Consequently, we made the decision to keep all paragraphs which had at least 2 declarative sentences (i.e. sentences which end in a period), and remove the rest. From our observations, even though some potentially useful text may have been lost in the process, using the aforementioned strategy resulted in more homogeneous samples which were less prone to causing overfitting. We also chose to ignore the files which were left empty using this approach, which led to one of the labels only having 447 samples left. We thus proceeded to balance the dataset, randomly selecting 447 samples from each difficulty level.

	WeeBit	WeeBit-NoSpecialChar	WeeBit-NoSpecialChar Filtered & Balanced
Overall	10486	10439	2235
Ages 7-8	629	628	447
Ages 8-9	789	789	447
Ages 9-10	807	807	447
Ages 11-14	646	646	447
Ages 15-16	7615	7569	447

Table 1: Number of samples per difficulty level for each of WeeBit's pre-processing stages

3.3 Metrics

In order to assess each model's performance, we decided to use two metrics: **accuracy** (number of correct predictions divided by total number of samples) and **RMSE** (root mean squared error). Even though the latter metric is not used as

universally as the former, it has been found to be particularly effective in "revealing performance differences" [4], a quality which is of great interest to our research. RMSE, defined by Equation 1, works by computing a value which corresponds to how far away the predictions are from the actual labels. A bigger "distance" is thus translated into a larger RMSE value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (1)$$

3.4 Exploration Protocol

The first step in our exploration was training the models on the modified WeeBit dataset presented above, using a selection of 80% of the samples. The remaining 20% made up the test set. The same train-test split was used for all models. BERT, RoBERTa, BART, and GPT-2 can be easily fine-tuned for performing readability assessment by using the Python libraries *transformers* and *datasets*, while ReadNet's code is freely available online in its entirety. We chose to train ReadNet across 100 epochs.

Having obtained the models' predicted label for every sample in the test set, we had to check whether the differences in performance between the five text difficulty classes were significant for every model. Since we needed to find the building blocks of our two metrics, accuracy and RMSE, we conducted two one-way ANOVA tests. That is because running the models returns a predicted label and an actual label for every sample, which have to be converted into values that ANOVA can measure. Accuracy is an aggregate of truth values (whether each prediction is correct or wrong), while the RMSE is based on a simple error calculation (the modulus of the actual label minus the predicted label). We thus used these as values for the ANOVA tests, with each label as a group. We considered all p values which were lower than 0.05 to be statistically significant.

Using the predicted values and the actual labels, we proceeded to calculate the accuracy and RMSE of the models overall and for each difficulty label.

Comparing the results obtained by using the models to the values yielded by the Flesch-Kincaid grade level formula was not a straightforward task, since the WeeBit corpus classifies texts by age, not by grade level. Therefore, in order to be able to compare the Flesch-Kincaid grade level formula with the models we studied, we had to first find the likely grade levels for each age label. This classification has been calculated starting from the kindergarten entrance age of 5 which is observed by most states in the USA [14]:

- 7-8 years old = 2nd or 3rd grade
- 8-9 years old = 3rd or 4th grade
- 9-10 years old = 4th or 5th grade
- 11-14 years old = 6th grade to 9th grade
- 15-16 years old = 10th or 11th grade

Having made this classification, the accuracy and RMSE of the Flesch-Kincaid grade level formula were evaluated on WeeBit data by comparing the value returned by the formula to the grade interval corresponding to the sample's label. If the value was inside the interval, then the formula's prediction was considered correct, with an error of 0 by definition. If however, the value was outside the interval, then the prediction was considered incorrect, with an error of $Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the end of the interval closest to the predicted value. For example, for the label *9-10 years old* and a predicted grade level of 11, the error is $Y_i - \hat{Y}_i = 11 - 5 = 6$, since the end of the interval *4th grade to 5th grade* closest to the predicted value of 11 is 5.

4 Results

In this section, we present and discuss the empirical exploration's overall results, we conduct an in-depth analysis of the models' performance per text difficulty class, and we contrast the models with the baseline.

4.1 Overall Performance

For determining the overall performance of the models, we compare and contrast the correctness of their readability estimations on the entire dataset. For evaluating this correctness, we consider two metrics, accuracy and RMSE, which tell us how many predictions were right, and how far were the predictions from the correct result, respectively.

We first examine the overall performance of the fine-tuned **BERT**, **RoBERTa**, and **BART** models, since in their case the results were close. As captured in Tables 2 and 3, BERT ranked third and first, RoBERTa ranked first and second, and BART ranked second and third in terms of accuracy and RMSE, respectively. Even though RoBERTa has a slight edge on the other architectures, it is still difficult to ascertain which one of these models performed best overall.

The fine-tuned **GPT-2**'s performance was on par with the ones of the three models analyzed above, even though it ranked fourth for both accuracy and RMSE. The difference between its performance and the one of the accuracy and RMSE leaders is not substantial, 0.036 for the former and 0.198 for the latter.

The readability model **ReadNet** exhibited a much poorer performance than the other four models. Its accuracy and RMSE lagged behind the fine-tuned GPT-2's, with differences of 0.405 and 0.729, respectively.

	BERT	RoBERTa	BART	GPT-2	ReadNet	Flesch-Kincaid
Overall	0.789	**0.796**	0.791	0.760	0.364	0.240
Ages 7-8	0.833	**0.844**	0.822	0.833	0.000	0.200
Ages 8-9	**0.778**	0.711	0.711	0.567	0.078	0.333
Ages 9-10	0.644	0.722	0.711	0.733	**0.833**	0.133
Ages 11-14	0.867	0.911	**0.922**	0.833	0.800	0.467
Ages 15-16	0.822	0.789	0.789	**0.833**	0.111	0.067

Table 2: Accuracy on WeeBit, by model and text difficulty level

4.2 Performance across Age Levels

Before analyzing the models' scores per each text difficulty category, it is useful to stress that WeeBit does not partition its samples consistently. The first three labels, 7-8, 8-9, and 9-10 have some target age overlap, so samples belonging to consecutive groups among the three are likely to have a high

	BERT	RoBERTa	BART	GPT-2	ReadNet	Flesch-Kincaid
Overall	**0.485**	0.527	0.542	0.683	1.412	2.296
Ages 7-8	**0.408**	0.527	0.548	0.775	2.541	1.542
Ages 8-9	**0.558**	0.641	0.615	0.782	1.374	1.359
Ages 9-10	0.624	0.558	0.568	0.683	**0.408**	1.555
Ages 11-14	0.365	0.422	**0.279**	0.506	0.447	1.721
Ages 15-16	**0.422**	0.459	0.624	0.632	1.121	4.092

Table 3: RMSE on WeeBit, by model and text difficulty level

degree of similarity. Moreover, the category 11-14 is way broader than the others, covering 4 ages instead of 2. This may also affect performance, just as it did for the Flesch-Kincaid grade level formula, which achieved a score way higher than its average on this label. This is due to a higher amount of grade level predictions being considered correct for the wider age interval.

As can be observed in Table 4, there exist significant differences of performance between text difficulty levels for every model, with p values ranging from 0 to 0.03.

Based solely on its achieved accuracy scores, the fine-tuned **BERT** model had an interesting behavior. It performed best on the first and last two age categories, reaching accuracy scores higher than 0.8 for all of them. Surprisingly, the only subdivision on which BERT surpassed its peers was 8-9, even though its accuracy on this category was rather low, at 0.778. On the 9-10 category however, BERT grossly underperformed. In terms of RMSE, the BERT-based model outclassed its counterparts on three text difficulty subdivisions, and showcased the best overall score. This means that, on average, its predictions were the closest to the actual labels, even in some cases when other models achieved higher accuracy scores.

The fine-tuned **RoBERTa**'s impressive overall accuracy score looks somewhat unjustified when its performance is analyzed per label. It did reach an accuracy of 0.844 on 7-8, which was the best result achieved for this category, and its accuracy was higher than 0.9 on 11-14. However, its performance on the other three labels, though not bad, was rather unimpressive, and its RMSE scores completely lagged behind BERT's.

The **BART**-based model had a similar performance to RoBERTa's. It surpassed its peers on the 11-14 age bracket on both accuracy and RMSE, and also performed well on the 7-8 subdivision in terms of accuracy. Its performance on the rest of the labels though was unremarkable.

In some ways, the fine-tuned **GPT-2** model behaved similarly to BERT. It also reached accuracies higher than 0.8 on the first and last two age categories (achieving the highest score for 15-16, 0.833), and it underperformed on one label, 8-9 in this case. Its RMSE scores though were not comparable to neither BERT's nor RoBERTa's on any subdivision.

As for **ReadNet**, this model is classifying most samples in just two categories, ages 9-10 and 11-14. This leads to high accuracies and low RMSEs for those 2 labels (with the highest accuracy of 0.833 among all models on 11-14, class on which ReadNet has the lowest RMSE as well), but also to very low scores for the rest of the labels. Its accuracy on the 7-8 category for example is a "perfect" 0.

	BERT	RoBERTa	BART	GPT-2	ReadNet
correct/incorrect	0.002	0.003	0.002	0.000	0.000
error	0.002	0.030	0.009	0.001	0.000

Table 4: ANOVA p-values, by model and value type

4.3 Comparison to Baseline

Our baseline of choice consists of one of the most popular formulas in readability assessment, the Flesch-Kincaid grade level formula, whose performance is illustrated in Figures 2 and 3. It achieves an overall accuracy of 0.240, which is just a little better than chance (choosing randomly between the five difficulty levels), and its RMSE of 2.296 is significantly higher than the scores demonstrated by the transformer models. The fine-tuned BERT, RoBERTa, BART, and GPT-2 models reach considerably superior scores on all text difficulty classes, for both accuracy and RMSE, so we can conclude that these four models are viable tools which can be used for assessing the readability level of texts, regardless of their difficulty levels. ReadNet however underperforms the formula on the lower text difficulty subdivisions, achieving worse accuracy and RMSE scores. This prompts us to infer that ReadNet is not a viable tool which can be used for assessing the readability level of texts targeted towards children aged 9 or lower.

5 Responsible Research

Since reproducibility is one of our main concerns, the code used for conducting the experiments presented in this paper can be found at *https://github.com/dsachelarie/transformers-readability-assessment*, with full instructions on how to run it in the *README.md* file.

Our use of WeeBit has been authorized by Prof. Dr. Detmar Meurers, co-author of the paper which introduced the corpus [17]. Since we do not have the right to share the corpus with third parties, it has not been provided in the repository. We recommend anyone who needs access to the corpus to contact Prof. Dr. Detmar Meurers at the following email address (active at the time of writing): *dm@sfs.uni-tuebingen.de*.

The ReadNet code we used was also excluded from the repository, since most of it was not created by us. It consists of the code which is freely available at *https://github.com/vdefont/readnet*, with some additions inspired from *https://github.com/Nobert1/information-retrieval*. By adapting the code provided in the aforementioned repositories, as well as consulting the paper which introduced the model [13], ReadNet can be easily integrated into our repository's code.

6 Discussion

The results which were presented in the previous chapter are sufficient for answering the research question. The accuracy and RMSE scores per difficulty level offer us an extensive picture of each model's strengths and weaknesses, when their performance on texts targeted towards school age children is considered. For a more high-level discussion, it is helpful to make a distinction between lower age (7-10) and higher age (11-16) texts, an idea which was inspired

from WeeBit's composition, since this corpus was formed by putting together information from Weekly Reader (ages 7-10) and BBC-Bitesize (ages 11-16) [17]. According to *https://www.bbc.co.uk/bitesize*, the official website of BBC-Bitesize, the higher age interval we set actually corresponds to the secondary school level, while the lower age interval only contains texts targeted towards the primary school level. We will concentrate on the strengths and weaknesses of the pre-trained models, since, in our view, ReadNet's pronounced lack of consistency is a very strong weakness, which makes assessing any of this model's strengths impossible.

The two models which achieve good scores consistently on higher age texts are **BERT** and **GPT-2**. Both models however underperform on one lower age difficulty subdivision. There is actually no model which achieves good scores on all lower age texts, but **RoBERTa** and **BART** both reach decent results. We can thus observe that where two of the pre-trained models are strong, the other two are weak, and vice versa.

Another strength of the fine-tuned **RoBERTa** and **BART** models is their overall consistency. They never actually underperform, achieving decent results on all labels, which makes them especially reliable when texts of many difficulty levels are being evaluated.

As for the fine-tuned **BERT** model, its main strength lies in its propensity to achieve lower RMSE scores than its peers. This means that it should be the tool of choice in situations when a low error is more desirable than a high accuracy.

The fine-tuned **GPT-2** model proved to be suitable for readability assessment, since its performance measured in terms of accuracy and RMSE was close to what the other transformer models achieved on every text difficulty subdivision. A more recent transformer from the GPT family may potentially reach even better results.

Our overall accuracy results for the fine-tuned **BERT**, **RoBERTa**, and **BART** models are significantly lower than what was found in previous research. The first related paper we presented [9] claimed accuracies of 0.893, 0.900, and 0.889 for readability models based on BERT, RoBERTa, and BART, respectively. Our models performed worse though, reaching overall accuracies of 0.789, 0.796, and 0.791, respectively. We believe that this significant difference in accuracy scores may be due to our decision to remove unwanted text from the samples, since we wanted to prevent the model from correlating difficulty levels with certain snippets of "garbage" text. The paper [9] never actually mentions any unwanted text removal, so it is likely that such text was still present in the samples when the models were trained and tested. Our conviction that this is the cause of the difference is also supported by our experience with pre-processing WeeBit, since in earlier experiments, we achieved some very similar overall results to what was found by [9], due to only attempting limited "garbage" text removal. That being said, our work and the research presented by [9] had one similarity in the results which were reached: RoBERTa performed best in terms of overall accuracy, being closely followed by BERT and BART.

ReadNet did not match the performance that was claimed in the paper that introduced it [13], with its overall accuracy being only slightly better than the one achieved by the Flesch-Kincaid grade level formula, the experiment's baseline. The paper [13] does not specify any "garbage" text removal, so the high accuracy results may have been influenced by correlations between some unwanted text snippets and the sample labels. That being said, it is possible that we could've achieved a better ReadNet performance if we trained it on more than 100 epochs, but even with this modification, it is unlikely that this model would have been able to reach the same results as its counterparts.

7 Conclusions and Future Work

The aim of this empirical exploration was to find the strengths and weaknesses of several transformer models used for readability assessment. After evaluating the performance of four models based on the BERT, RoBERTa, BART, and GPT-2 architectures, as well as of the standalone readability model ReadNet, we demonstrated that there are indeed differences in each model's performance on the many text difficulty levels. Considering each transformer's strengths and weaknesses, we deduce that a RoBERTa-based model is the overall best performing transformer option for readability assessment, when the text being evaluated is targeted towards school age children. A fine-tuned RoBERTa is also the best performing option when only lower age (7-10) texts are considered, while a BERT-based model is the most suitable for higher age (11-16) texts.

A limitation of our research, which could be addressed in future work, arises from ReadNet's poor performance. This model makes predictions by returning a real number between the minimum and maximum index of the available labels, instead of offering the predicted label or a list of probabilities. This behavior prompts us to believe that this model may be more suitable for binary classification tasks. Future research could find ways to boost its performance when ran on datasets with many labels.

The main limitation of our research though was our use of only one corpus. We ran our models solely on WeeBit, since we were not able to get access to NewsEla (*https://newsela.com/data/*), another large and granular corpus which would have provided us with the means to generalize our findings. Future research will hopefully fill this gap by researching the behaviour of readability models not only on WeeBit and NewsEla, but on several other corpora as well, therefore achieving full confidence in the conclusions which will thus be reached.

Through our findings, we hope that we have opened several directions for future research. We concentrated our work on how different transformer models perform on various text difficulty levels, but there are many situations in which the models could exhibit variability in their behavior: differing text lengths, text subjects, etc. However, even if no additional related research is conducted, our work is nevertheless relevant. We have found a few rough strengths and weaknesses of some of the best performing transformer models used in readability assessment, which should make choosing between several models for specific readability assessment tasks a much easier endeavour.

References

[1] Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. Supercalifragilisticexpialidocious: Why using the "right" readability formula in children′s web search matters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 3–18, 4 2022.

[2] Xavier Amatriain, Ananth Sankar, Jie Bing, Praveen Kumar Bodigutla, Timothy J. Hazen, and Michaeel Kazi. Transformer models: an introduction and catalog, 2023.

[3] Ion Madrazo Azpiazu and Maria Soledad Pera. An analysis of transfer learning methods for multilingual readability assessment. *UMAP 2020 Adjunct - Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 95–100, 7 2020.

[4] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 6 2014.

[5] Edgar Dale and Jeanne S. Chall. The concept of readability. *Elementary English*, 26(1):19–26, 1949.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[7] Joseph Eastwood, Brent Snook, and Kirk Luther. Measuring the reading complexity and oral comprehension of Canadian youth waiver forms. *Crime and Delinquency*, 61(6):798–828, 1 2015.

[8] Manas Gaur, Keyur Faldu, and Amit Sheth. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, 25(1):51–59, 2021.

[9] Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 10669–10686, 11 2021.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[12] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, 4 2021.

[13] Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. Readnet: A hierarchical transformer framework for web article readability analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12035:33–49, 4 2020.

[14] National Center for Education Statistics. Table 1.3. Types of state and district requirements for kindergarten entrance and attendance, by state: 2020. https://nces.ed.gov/programs/statereform/tab1_3-2020.asp. Accessed: 19-06-2023.

[15] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[16] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[17] Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7), Association for Computational Linguistics*, 2012.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.