Performance of Transformer Models in Readability Assessment

David Sachelarie[†*] Responsible Professor: Maria Soledad Pera[*]

[*]EEMCS, Delft University of Technology, The Netherlands, [†]d.sachelarie@student.tudelft.nl

Context

- **Readability** = how well can a text be understood by its readers;
- **Transformer** models = neural networks which rely on attention mechanisms, thus not needing recurrence and convolutions [1];
- Transformer models generally achieve better results than most other readability assessment tools [2;3];
- We chose five models: the standalone readability model **ReadNet**, and models based on the **BERT**, **RoBERTa**, **BART**, and **GPT-2** architectures.

Motivation

- Understanding how well transformer models perform in different situations is essential for deciding when to use each readability tool in order to get optimal results;
- It is unclear which classes of text difficulty each model performs best and worst on;
- There are no empirical works which compare and contrast **ReadNet** to other transformers;
- The **GPT** family of transformers has never been tested in readability assessment;
- Research question:

What are the strengths and weaknesses of various transformer models used for readability assessment?

Experiment

I Fine-tuned **BERT**, **RoBERTa**, **BART**, and **GPT-2** for readability assessment;

II Pre-processed and downsampled **WeeBit**;

III Trained the **fine-tuned models** and **ReadNet** on **WeeBit**;

IV Evaluated the performance of the models, as well as of the baseline, the **Flesch-Kincaid grade level formula**, using **accuracy** and **RMSE**.

Findings

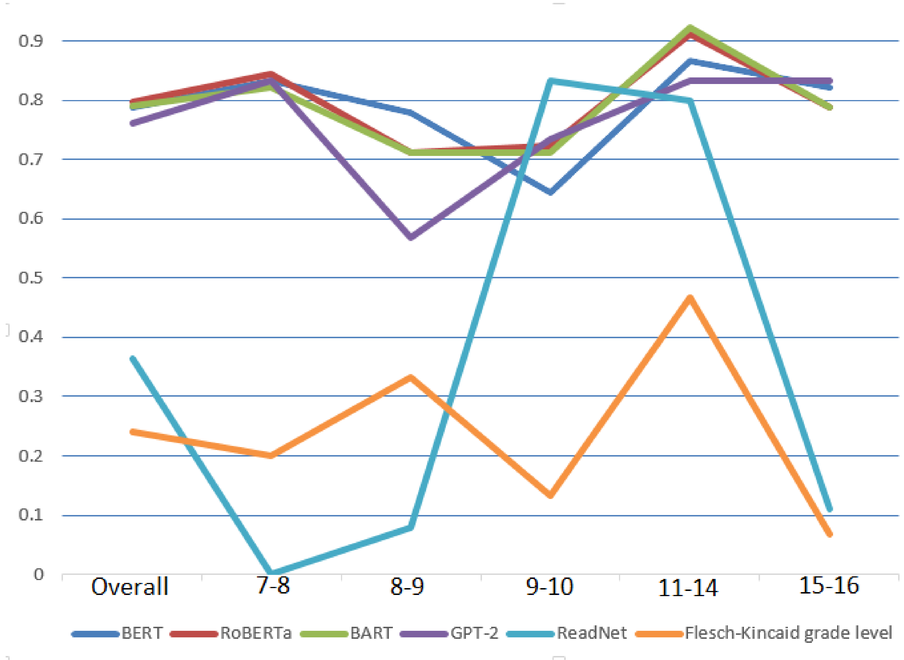

Figure 1: Accuracy per age bracket

- The fine-tuned **BERT**, **RoBERTa**, **BART**, and **GPT-2** models are better than the baseline on all labels;
- **RoBERTa** and **BART** are the best options for lower age texts (7-10), and **BERT** and **GPT-2** for higher age texts (11-16);
- **RoBERTa** and **BART**'s consistency makes them best overall;
- **GPT-2** is suitable for readability assessment;
- **ReadNet**'s performance was way poorer than in the paper that introduced it [2];
- Lower accuracies were achieved by **BERT**, **RoBERTa**, and **BART** than what previous research claims [3];
- **BERT**'s predictions tended to be the closes to the actual labels, even in some cases in which other models achieved higher accuracy scores.

Conclusions and Limitations

- All models exhibited certain strengths and weaknesses;
- ReadNet may be more suitable for binary classification tasks, so future research could find ways to boost its performance when ran on datasets with many labels;
- We concentrated on one corpus, future research will hopefully analyze the models' performance on several corpora.

References

[1] 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.

[2] 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.

[3] 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.

TU Delft