**‘The Automated Poets Society’:**

**Using Deep Learning to Automatically Identify**

**High School English Texts**

Prepared by

Matt Bilton

22/09/2021

**Executive Summary**

This report outlines the use of a deep learning model to identify appropriate texts for High School Students within the Top 100 most popular texts on the Project Gutenberg repository of books.

**Introduction**

When it comes to developing the reading skills of High School English students, it’s critical that texts of the appropriate difficulty level are assigned: texts which are too easy to understand do not challenge student’s reading comprehension and, conversely, students are not likely to critically engage with texts which they find too difficult to understand.

Despite the educational importance of assigning ‘difficulty ratings’ to texts, this is a non-trivial task for three reasons:

1. Traditional methods
2. Commercially available software which is capable of this task are typically ‘non-transparent’ in the sense that their code is not open source – this means that

This these points in mind, there is clearly a need for an efficient way to evaluate the difficulty level of pieces of text with human-level accuracies.

Thankfully, deep learning algorithms have seen a great rise in popularity over the last decade, primarily due to their ability to rival, and even surpass, human level performance in a variety of tasks, including tasks involving language understanding.s

In this report, we’ll describe the

We’ll then describe the application of this model to a collection of the Top 100 most popular texts on the Project Gutenberg public repository; from these predictions,

**Overview of Methodology**

Deep Learning Model

Project Gutenberg Texts Scraping

Top 100 texts most popular texts were scraped. Each text was ‘sliced’ into smaller text chunks of roughly 250 words; the readability score of each of these chunks was then computed. This resulted in a collection of multiple readability values being computed for each scraped text (i.e. there is a readability value for each text chunk within the text). This meant that an **average** readability score could be computed for each text; additionally, the **variance** in the readability scores for each text could also be computed.

Text Recommendation Procedure

For the purposes of assigning texts to particular year levels, each text was classified as being either *‘Easy’*, *‘Medium’*, or *‘Hard’* on the basis of their readability score:

* *Easy* texts are those with an *average* readability score **above -1.0**
* *Medium* texts are those with an *average* readability score which is **between -1.0 and -2.0**
* *Hard* texts are those texts with an *average* readability score which is **less than -2.0**

Obviously, harder texts should be prescribed to older students, whereas easier texts should be given to younger high school students; the exact year levels which correspond to each difficulty band will need to be decided upon by individual teachers.

It was assumed that teachers would prefer books with more *consistent difficulty levels* over those books with highly variable difficulty levels. Put simply, this simply means that a book which frequently changes between very hard passages of text and very easy passages of text is less useful than a book whose passages are all roughly of the same difficulty. Under this assumption, **top five least variable texts within each difficulty band were chosen** as the ‘best texts’ to prescribe to students. A visual summary of the procedure used to choose these texts is shown in Figure X.

**Recommended Texts**

Based on the procedures described in the previous section, the following set of fifteen texts (i.e. five for each difficulty level):

|  |  |
| --- | --- |
| **Difficulty** | **Texts** |
| Easy (greater than -1.0 readability) | 1. *The Wonderful Wizard of Oz* 2. *Old Granny Fox* 3. *The Metamorphisis* 4. *Grimms’ Fairy Tales* 5. *Alice’s Adventures in Wonderland* |
| Medium (between -1.0 and -2.0 readability) | 1. *A Doll’s House* 2. *Treasure Island* 3. *The Call of the Wild* 4. *Pygmalion* 5. *The Importance of being Earnest* |
| Hard (less than -2.0 readability) | 1. *Leviathan* 2. *The Tragic Tale of Doctor Faustus* 3. *The Poetics of Aristotle* 4. *The Confessions of St Augustine* 5. *Beyond Good and Evil* |

Perhaps pleasingly, a lot of ‘classic’ texts have managed to find their way , including Lewis Caroll’s *Alice’s Adventures in Wonderland*, Robert Louis Stevenson’s *Treasure Island*, and Nietzsche’s *Beyond Good and Evil*.

**Limitations**

There are three key limitations to note about the approach we’ve described in this report:

1. The model was trained to identify complexity in language and not necessarily complexity of themes, which is just as important. To see why this might matter, consider a
2. Only looked at texts – this repository only includes texts which are publicly available and are not protected by copyright. Because of this fact, the Gutenberg Project almost exclusively holds *old* texts, since the copyright on these books has long since expired. This means that our \_ was unable to recommend any modern texts which may be more appropriate for students.
3. Unfortunately, there was not enough time to perform an interpretability analysis on the outputs of the deep learning model – this means that the model we’ve trained here is effectively a ‘black box’ in the sense that we don’t actually completely understand how it makes its predictions. Future work could perhaps

**Conclusions**

* A deep learning-based model called BERT was trained to predict the ‘readability score’ of excerpts of text.
* Using this, a series of five texts for each difficulty level was recommended
* The main limitation of the methodology

**Executive Summary**

* Brief overview of the report
* Give reader a preview of the content
* Present key points of the report
* Explain why you wrote the report, emphasize conclusion and recommendations. Only include the essential information to support the conclusion
* 1-2 paragraphs

**Table of Contents**

* List headings and subheadings

**Introduction**

* Describe the context, background, and problem of the business case
* Outline the report scope
* Briefly comment on limitations
* 2-3 paragraphs

**Analysis**

* Describe the methodology/model
* Discuss results

**Conclusion**

* Finish up the report starting from the objective to the recommendation

**Appendices**

* Place all material/data important to your analysis that is referred to in your discussion
* Title material APPENDIX A, APPENDIX B, etc., and in order of which they appear in the report