‘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 development of a deep learning-based tool to quantify the difficult of texts for the purpose of assigning appropriately difficult texts to High School English students. To illustrate the practical usability of our deep learning model, we use it to quantify thee difficult of the top 100 most popular texts in the Project Gutenberg repository. Based on these difficulty quantifications, we then recommend five different texts across three different difficulty bands to be given to students.

Contents

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 used to evaluate text difficulty typically rely on very simple heuristics (e.g. more complex texts have longer sentences), which can sometime produce inaccurate results for certain texts (e.g. long sentences of very simple text)
2. Text difficulty scores based on asking a sample size of readers to 'rank' the difficulty of a text are labor intensive – they normally require hundreds of people to read a particular text for statistically accurate difficulty score to be produced.
3. 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 teachers and programmers just need to trust that the commercially software is working as promised, with no way to easily verify whether this is true or not.

With 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.

Over the last decade, deep learning algorithms have seen a great rise in popularity. These 'deep learning algorithms' work by learning from example in a process called *model training* – for the purposes of a language understanding task, such as translating text from one language to another, a deep learning model would be provided with hundreds of millions of examples of translated sentences from which it could learn.

The increased usage of deep learning has primarily due to the ability of these methods to rival, and even surpass, human level performance in a variety of tasks, including tasks involving language understanding. Consequently, it seems that such deep learning algorithms may provide an attractive way to easily and accurately quantify the difficulty of pieces of text.

For the remainder of this report, we'll describe the training of a deep learning model to quantify the difficulty of excerpts of text. To illustrate the practical usefulness of this model, we then use our trained model to ascribe a difficulty score to the Top 100 most popular texts in the Project Gutenberg repository. From these difficulty ratings, we then recommended a series of texts for different year levels at high school.

Overview of Methodology

Deep Learning Model Training

We chose to train the BERT deep learning model for the purposes of creating a text difficulty evaluation tool. BERT is an incredibly effective, publicly available language processing model which has been trained on billions of pieces of text. Because of the sheer amount of data BERT has been trained on, it has a 'statistical understanding' of language – given one half of a sentence, it's likely able to guess what will be written in the next half of that sentence.

Our BERT model was 'finetuned' on a dataset made available by the CommonLit organization through Kaggle (which is simply an online platform to host data science competitions). This dataset consisted of a set of text extracts along with a numerical difficulty value associated with that text, which was found by asking hundreds of English teachers to rank the difficulty of each text extract.

Project Gutenberg Texts

As was previously mentioned, the entirety of the top 100 texts listed on Project Gutenberg were extracted and fed as inputs into our trained deep learning model. More specifically, each text was 'sliced' into smaller text chunks of roughly 250 words; the difficulty score of each of these chunks was then computed. This resulted in a collection of multiple difficulty values being computed for each scraped text (i.e. there is a difficulty value for each text chunk within each book). 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 1 below.

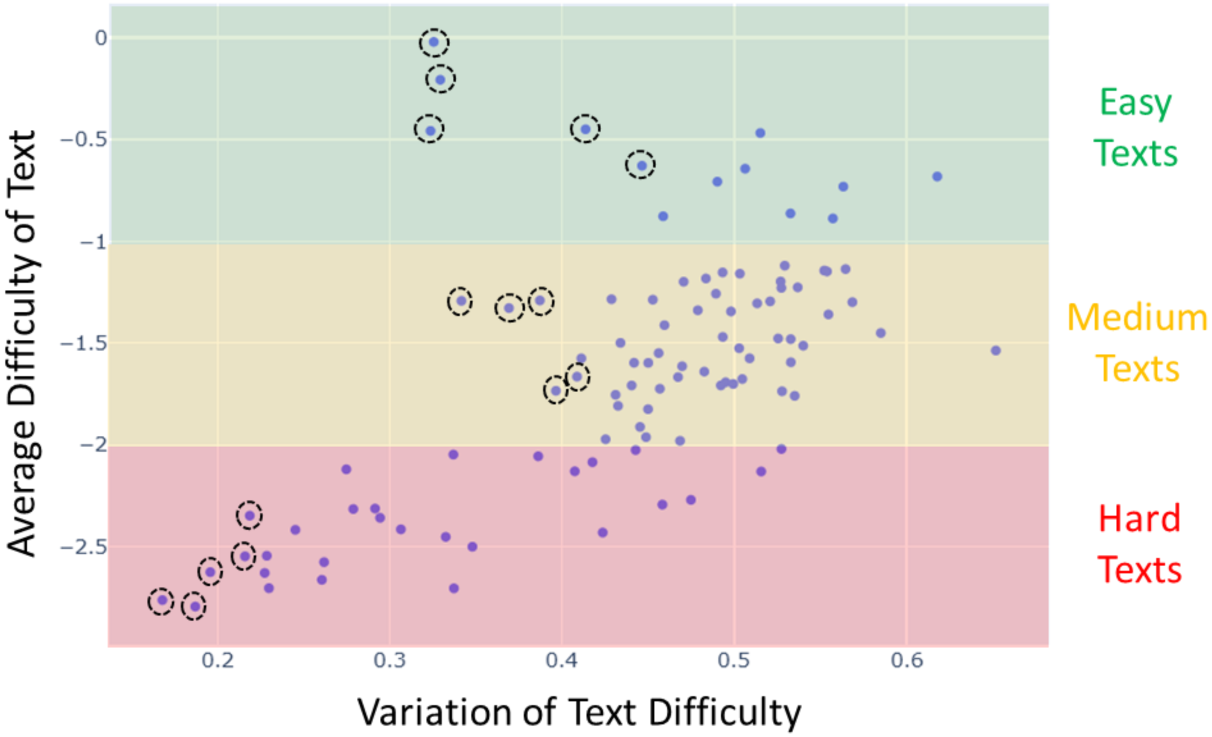

Figure 1: The average difficulty of each Gutenberg text plotted against the variation of the text difficulty within that text. The difficulty score cutoffs for each are shown, and the least variable, chosen texts within each band are circled.

Recommended Texts

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

Table 1: Project Gutenberg texts recommended for each reading 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 into our recommendations, 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 two key limitations to note about the approach we've described in this report:

1. Our BERT model was finetuned on a relatively small dataset of 2000 examples; the performance of our model may be impaired if this dataset was biased or non-reflective of the greater 'population' of texts which exist.
2. Our text recommendations were informed only looked at texts available within the Project Gutenberg repository. Generally speaking, this repository only includes texts which are publicly available and, thus, are no longer 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 we were unable to apply our model to more modern texts, which may have been more appropriate to recommend to 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.

Conclusions

- A deep learning-based model called BERT was trained to predict the 'readability score' of excerpts of text.
- This trained BERT model was then applied to quantify the difficulty of the top 100 most popular texts within the Project Gutenberg repository
- Based on these difficulty quantifications, a series of five texts for each difficulty level was recommended from these Gutenberg texts.

- The main limitations of the methods described in this report include a relatively small training data set, only looking at Project Gutenberg texts for recommendations, and not performing interpretability analysis to understand how our BERT network is making its difficulty predictions.