SMU Data Science Review

Volume 5 | Number 3 Article 7

2021

Reading Level Identification Using Natural Language Processing Techniques

William Arnost
Southern Methodist University, warnost@gmail.com

Ellen Lull
Southern Methodist University, ellenlull1@gmail.com

Joseph Schueder
Southern Methodist University, jjschued@gmail.com

Joseph Engler
Collins Aerospace, joseph.engler@collins.com

Recommended Citation

Arnost, William; Lull, Ellen; Schueder, Joseph; and Engler, Joseph (2021) "Reading Level Identification Using Natural Language Processing Techniques," *SMU Data Science Review*: Vol. 5 : No. 3 , Article 7. Available at: https://scholar.smu.edu/datasciencereview/vol5/iss3/7

Reading Level Identification Using Natural Language Processing Techniques

William Arnost[1], warnost@gmail.com
Ellen Lull[1], ellenlull1@gmail.com
Joseph Schueder[1,2,] jjschued@gmail.com
Dr. Joseph Engler[2,] joseph.engler@collins.com
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Collins Aerospace, Four Coliseum Centre, 2730 West Tyvola Road, Charlotte NC 28217-4578 USA

Abstract

This paper investigates using the Bidirectional Encoder Representations from Transformers (BERT) algorithm and lexical-syntactic features to measure readability. Readability is important in many disciplines, for functions such as selecting passages for school children, assessing the complexity of publications, and writing documentation. Text at an appropriate reading level will help make communication clear and effective. Readability is primarily measured using well-established statistical methods. Recent advances in Natural Language Processing (NLP) have had mixed success incorporating higher-level text features in a way that consistently beats established metrics. This paper contributes a readability method using a modern transformer technique and compares the results to established metrics.

This paper finds that the combination of BERT and readability metrics provide a significant improvement in estimation of readability as defined by Crossley et al. [1]. The BERT+Readability model has a root mean square error (RMSE) of 0.30 compared to a BERT only model with RMSE of 0.44. This finding offers an alternative to basic statistical measures currently offered by most word processing software.

1 Introduction

Literacy is a major factor impacting various aspects of an individual's quality of life and even affecting local, national, and world economies. For an individual, their ability to read is the primary building block for learning other subject areas. For example, proficiently reading material on science, geography, or math will allow the student to learn much more efficiently and deeply than hearing a lecture or watching a video on a given topic. For communities, having a citizenry that can read and learn creates opportunities for that community to grow culturally and economically. Reading skills are important to success in various aspects of life.

Reading skills can be difficult to build with average reading scores declining in the United States for grades 4, 8, and 12 [31][32]. Classifying texts into different reading levels can assist students in improving their reading skills. The theory is that a reader will improve by choosing a text slightly above their current reading level. This level is often called their "frustration level" as opposed to a "comfort level." Matching students with texts that are challenging but comprehensible can assist in attaining reading skills [35].

Current methods of determining reading levels for elementary school students are dated and often proprietary [1,26]. Many calculations can be completed on a text that can serve as a proxy for reading levels. Modern baseline research in this area started in the 1940s with the Flesch-Kincaid Grade Level Index model, which was based on proxies like unique vocabulary, syllables, and length of sentences [5]. Though Flesch-Kincaid was the first, numerous other calculations include the Dale-Chall and textual cohesion methods [20,23]. However, most of these subsequent models are criticized as lacking measurement of higher-level language constructs [6,17].

Proprietary models lack transparency for the formulas they use. One of the most widely used assessments is the Developmental Reading Assessment (DRA) offered by Pearson, the leading company in educational products in North America. The package includes both reading materials at appropriate age levels and assessments for grade levels K-8. The cost is published as 360 US Dollars [26]. This cost would be for each instructor. This could add up very quickly for schools that do not have adequate funding and for non-profit groups.

As iterations of changes to existing formulas were introduced, they all received criticism in some aspects of their methodology. There are numerous readability formulas in place today. Most of them have been criticized for not achieving the stated purpose of finding reading material appropriate for the learner. These formulas are also said to focus primarily on word length and the number of words in a sentence and lack other aspects like cohesion [17]. In the 1990s, Microsoft Corporation introduced a readability tool in its word processing software, Microsoft Word. The software uses the Flesch-Kincaid method and has the same set of criticisms of the Flesch-Kincaid model with the addition of technical nuances to ensure the calculation is complete [7].

In the last several years, new methods of NLP have come into existence. These methods tried to incorporate lexical diversity, text cohesion, and other methods [17]. The methods proved no more beneficial than the basic calculations. The most recent research in readability has been with so-called transformer models in the NLP branch of the machine learning domain. Results of these models have shown some improvement over the baseline formulas. However, improvements have been nominal. One example is text cohesion [28] models, which measure continuity between sentences and across bodies of text. However, these models did not achieve statistically significant improvements over existing methods.

A readability solution that is freely available to educators throughout the nation will enable those educators to find and add new reading material to curriculums independently of the paid services of today. In addition to being fee-based, today's paid services have a limited number of rated materials. Adding readings that are of interest to children based on their location, culture, and interests could create more options for students than are available today.

1.1 Assisting the CommonLit Organization

A new reading level ratability solution will help CommonLit, a non-profit organization dedicated to helping Title 1 students achieve reading proficiency. CommonLit is a non-profit that provides free reading and writing lessons to over 20 million teachers and students. Working with Georgia State University, CommonLit sponsored a Kaggle competition to improve reading ratability methods. Existing methods can be inaccessible to some teachers because of cost. CommonLit hopes to provide a modern solution at no cost. This will help teachers select appropriate materials for their students and help underserved students reach reading proficiency [1].

Since CommonLit is primarily dedicated to helping Title 1 students achieve proficiency, this work will have far-reaching impacts for the schools under that umbrella. Title 1 is a U.S. government-sponsored education funding program. The goal of Title 1 is to help underprivileged and underserved populations across the nation achieve a quality education. The school districts served by Title 1 typically have characteristics like extreme poverty, homelessness, and English as a second language speakers. The national report card in 2019 found that reading scores were declining in 4th grade, 8th grade, and among males in the lowest reading percentiles of 12th grade [32][33].

With CommonLit supplementing the national Title 1 program, school districts can free up money from reading to expand their curriculum on more education programs. Success with Title 1 schools has produced higher graduation rates and lower dropout rates previously unseen in the United States before the program [29].

If a more accurate, automated method of reading level determination can be created, it could be used by a wide variety of educational organizations in addition to CommonLit. In addition to the functional benefit of free, self-service readability ratings of educator-selected passages, readability formula solutions using machine or deep learning currently are the same or marginally better than the traditional simple statistics-based methods. Combining NLP with machine and deep learning models could create significantly more accurate estimates of reading levels.

2 Literature Review

Readability is a measure of how difficult it is to read and comprehend text. Readability is important for a variety of reasons. It can affect an audience's willingness to engage with a piece of material. It can affect your Google search rating or jeopardize a customer's ability to use your product competently. Several formulas have been created to assess readability. Early measures like Flesch-Kincaid and Dale-Chall use lexical-syntactic features to assess readability. Critics of these formulas often argue that the lack of higher-level concepts like required inference or textual organization makes these formulas incomplete. Chall and Dale (1995) [23] suggest that reading level formulas distinguish lower difficulty texts but lack the complexity to capture differences in higher difficulty texts.

Researchers have identified several key components to readability. Dale and Chall 1949 [20] suggest factors related to the subject matter and content being read, the interests of the individual reader, and the criterion specific to the formula being used. Reading formulas often capture lexical-syntactic features, like average sentence length or syllables per word. Other key features of text organization [24], disposition of the reader, textual cohesion [25], and inference load [21]. Studies show these features have important effects on readability.

There are many readability formulas available today to assess the reading level of text. These formulas use statistical features like word counts, syllable counts, and sentence length to determine difficulty. The "Flesch-Kincaid Grade Level" and "Flesch Reading Ease" are commonly used for readability [2]. The formula below shows the calculation of the Flesch-Kincaid grade level.

Flesch-Kincaid Grade Level [4, p 6] [5, p3]

$$FK\ Grade\ Level\ = 0.39\left(\frac{total\ words}{total\ sentences}\right)+11.8\left(\frac{total\ syllables}{total\ words}\right)-15.59$$

Flesch-Kincaid is described as an accurate measure for school text; Dale-Chall Readability Formula is focused on difficult words and sentence length to categorize 4th grade and over 10th-grade readability [23]. Forecast uses the number of single-syllable words but not to be used to assess primary age reading materials. Methods that are used for children's books in the United States include Flesch-Kincaid Grade Level from 1975, Fry (for children under age 10), CLI for high school and older, Dale-Chall for determining 4th and 10th-grade levels, SMOG for secondary grades, and Spache which is similar to Dale-Chall.

Another way readability is measured is by ranking documents using the Bradley-Terry model. This model compares text documents to each other in a pairwise manner [30]. A person ranks which document "wins" by being easier to read than another document.

Textual cohesion refers to how parts of a text relate to each other. You can measure how one sentence relates to the next, count connective words and sentence structures, among other measures [27]. Todirascu et al. [28] describe cohesion more formerly as "a property of text represented by explicit formal grammatical ties (discourse connectives) and lexical ties that signal how utterances or larger text parts are related to each other" [28, p. 988]. They describe several cohesion features, including anamorphic chains, coreference chains, lexical chains, and others. Of sixty-five measures tested, only six were significant, and the incremental improvement over their baseline model was small.

Machine learning methods have been used to improve reading level identification. With this approach, NLP tools are used to extract features that can be fed into a machine learning model, such as a Support Vector Machine, Regression, Random Forest, or other types of models. Crossley et al. [17] discuss extracting the features from documents using NLP tools to extract syntactic features and sentiment analysis, then using these features in regression models. Determining how to measure word complexity is a big challenge in creating the features in a machine learning model. Gong et al. [14] discuss research to find a measure of word ambiguity that might help define a new measure from Wordnet.

Additional machine learning research by Sarah E. Petersen and Mari Ostendorf [6] used a Support Vector Machine (SVM) model approach used on a corpus of data from the popular magazine, "Weekly Reader." They imagined that a teacher was looking for a text on the web suitable for a class. Their paper describes how they combined NLP methods, including n-gram language models, parsing, smoothing, to get features that can be used with a Support Vector Machine (SVM) model to calculate reading level on Weekly Reader text. They compared the results to Flesch-Kincaid and Lexile reading level methods to test data that was annotated by human reading level experts. They received a more favorable F-score with their model. The F-score measures a model's accuracy which is calculated using a combination of precision and recall, which are individual accuracy measurements. Precision is the number of correct positive predictions from the model divided by the total number of positive predictions, whether correct or not. The recall is the number of correct positive predictions divided by the number of correct positive predictions plus the false negatives, or the ones that should have been positive but were not.

Neural Network models have also been used as another machine learning technique. Maddela and Xu. [12] created a lexicon of 15,000 commonly used English words and created a Convolutional Neural Network (CNN) model to determine the complexity of the words. Something similar was done by Aroyehun et al., who discusses comparing a CNN model to a Feature Engineering Model [10].

Martinc et al. [16] looked at several different approaches to the readability problem. Topics that were covered were traditional readability metrics, a new novel neural network model. For the neural network models, both supervised and unsupervised methods were explored. According to this paper, prior analysis of neural network approaches compared to the traditional measures rarely performed better than the traditional approach. This research made modest gains with neural networks at most one to two percentage points different. Specific neural networks were superior to others, but this depended on genre, length of the document, and language—most of the time. Hierarchical Attention Network performed better on longer texts, while BERT models performed better on shorter texts and foreign languages. This was thought to be because BERT has a limit on the token size. Bidirectional LSTMs were also compared and sometimes did a reasonable job. One unique aspect of this study was that four different training sets were used. These training sets had different properties. Those different properties included the distribution of reading levels, languages, and length of the documents. The distribution on some of the documents lent to a theory that if a corpus had a higher count of lower, middle, or higher-level documents, there was correlation to outcomes.

Graph models represent data in terms of nodes and edges. There have been several papers written describing different learning techniques using the graph data structure, including one by Kipf et. al. describing a scalable graph learning method that overcomes the limitations of other graph methods [41]. The paper states the previous methods often included multistep pipelines that were difficult to optimize. Other works that operate directly on the graph must learn node specific weight matrices that limit their ability to scale to larger data. Kipf et al. introduce simplifications to these methods that increase the scalability of a graph convolutional network.

3 Methods

This project built a model using a modern transformer method and other NLP techniques to predict reading levels. The model used BERT for embeddings as well as classical features. Traditional reading level measures were used to build features that would represent the complexity of a word or sentence that would indicate reading level.

The algorithm was trained using a labeled data set. The training data was split into eighty percent training data and twenty percent testing. The target variable for training is a Bradley-Terry score. Bradley-Terry is a model where items (in this case) text documents are compared in a pairwise manner [30]. A person, the rater, ranks which one "wins" or "loses." In this problem, a document wins by being easier than another document. Approximately 10,000 texts were chosen and given to human raters to evaluate in the data gathering and preparation. The texts were given to the raters as pairs. The raters then chose which of the two texts was easier to read or understand and which one took less time [17, p. 9-10]. Using these comparisons, the Bradley-Terry Model estimates the probability that one text is harder than the other. The model uses maximum likelihood estimation to estimate a parameter for each document. The higher the absolute value of the score, the harder the text is to read. The model produces a likelihood score [17, p. 9-10]. Since it is human labeled, it will be subjective and may complicate the accuracy measurement.

3.1 Training Data Visualization:

Base training data was obtained from the Kaggle competition to improve reading ratability methods which CommonLit sponsored. This dataset was chosen because it has been classified with labels to identify the readability of the text that serves as a benchmark to measure new models against. For this project, the Kaggle dataset has been augmented by readability scores from the python readability package [33], which has calculated Flesch-Kincaid, Dale-Chall, Gunning Fog, and Coleman-Liau scores. The Gunning Fog Readability Formula is a formula that uses the number of words in a sentence and considers the number of words with three syllables or more. The Coleman–Liau index is a grade-level calculator that focuses on characters instead of syllables per word [34].

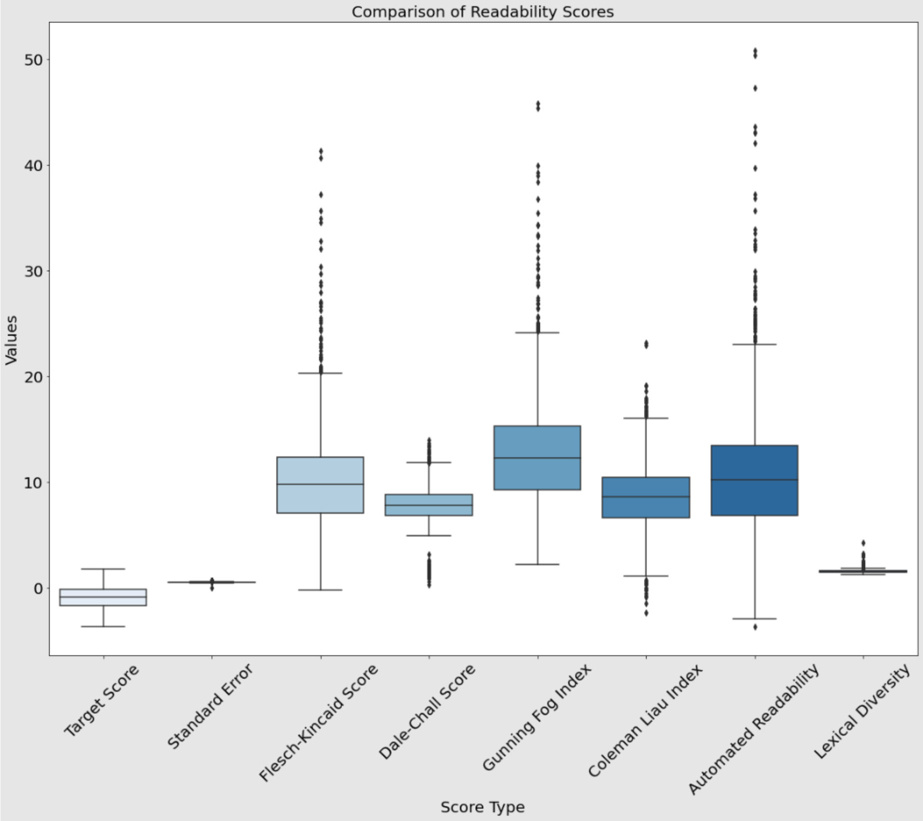

Figure 1: Boxplot of the features in the training dataset. It displays the distribution of the values of these features and gives a visualization of the outliers.

3.2 Description of BERT Model

BERT, or Bidirectional Encoder Representations from Transformers, is a NLP transformer model published by Google A.I. Language. Other models often read text left to right, but BERT reads in an entire sequence of text at once rather than from left to right [18]. Transformers are a class of attention mechanisms that create context relationships between tokens or words in a text [18,19]. Because transformers can process the data in any order, parallel processing can allow for copious amounts of training data. The attention mechanisms are the weighted connections between the output nodes and the input nodes, allowing higher priority on some connections than others to determine what is the most relevant context of the text.

The models have an encoder layer, shown on the left of figure 2, that maps the input sequence of the data and a decoder layer, shown on the right that maps the output sequence.

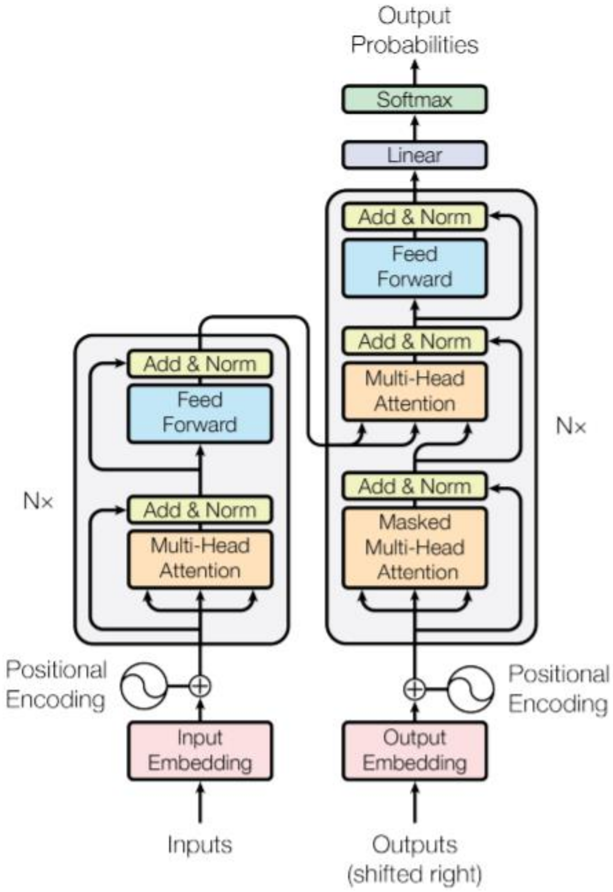

Figure 2. The Transformer – model architecture [18, p 3] "Attention is all you Need"

BERT uses many different subcomponents as seen in the BERT mountain (Figure 3), including encoding/decoding of text to numeric representations and masking. Masking is a pretraining that happens where some of the words are hidden from the model and the model is training to predict which words are masked. This helps the model create language context which is useful in a wide variety of NLP tasks. Finally, "attention" is given to words that seem more important than others via weighted averages. BERT is then "fine-tuned" for other natural language tasks via transfer learning [43].

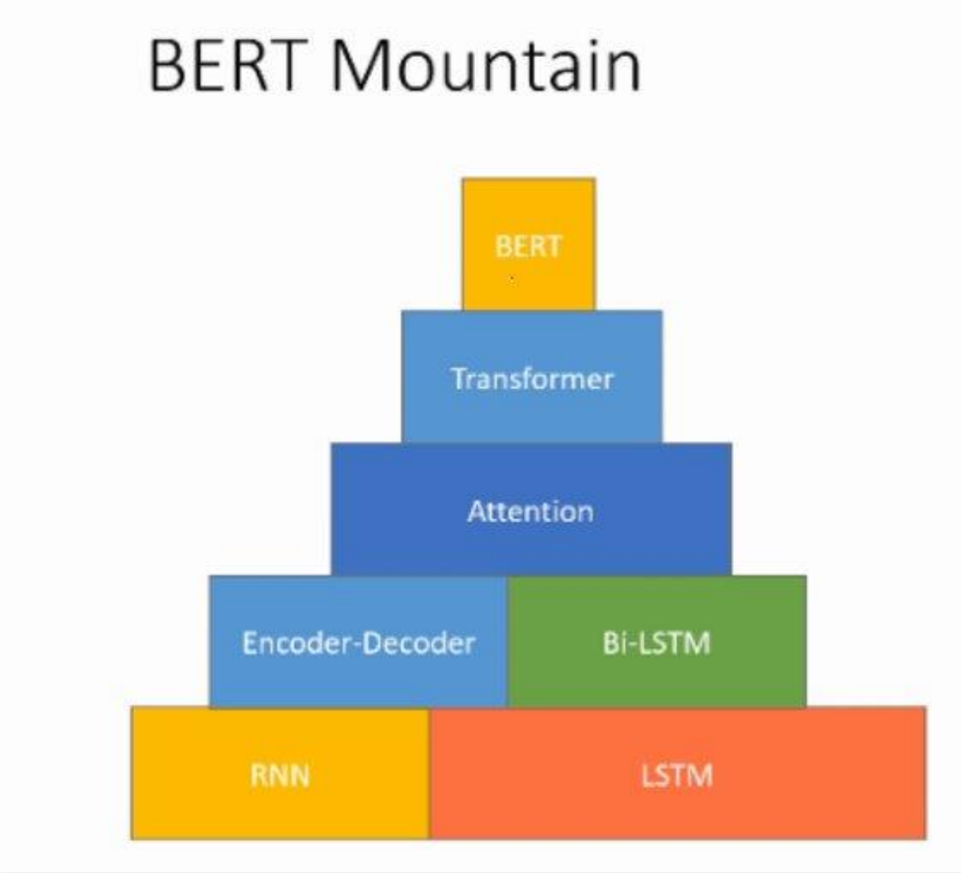

Figure 3. Bert Mountain [43] "Survey - BERT"

3.3 Graph Convolutional Network

BERT does an excellent job of capturing local context, but global concepts can still be challenging. Graph Convolutional networks can help map global relationships in addition to document specific ones [40]. A graph convolutional network was implemented using the Spektral package, which implements the graph convolutional architecture described by Kipf et al. except a linear activation on the output layer [41] was used. The graph is structured with the documents as nodes and edges defined by common linguistic elements. An edge between two documents occurs when they have at least five elements from the following list in common. This implementation used binning to make matching easier between nodes.

- Word Count
- Syllable Count
- Character Count
- Complex Word Count
- Vocab Size
- Lexical Diversity
- Noun Chunks
- Flesch Kincaid Score
- Dale Chall Score
- Gunning Fog Index
- Coleman Liau Index
- Automated Readability Index

The graph convolutional network has two graph convolution layers, the second layer has a linear activation function. The model was trained on 1000 documents, using 60/20/20 train / validation / test split.

4 Results

The dataset contains labels that are used for model training. The data was prepared using an 80/20 train/test split to create an unbiased accuracy metric for the proposed model. RMSE was the metric used to measure accuracy. The final test data was obtained from Dr. Scott Crossley after the close of the Kaggle competition to compare competition results to the results in this paper. This data was not used for model training.

This problem contains continuous numeric values as the target labels. Due to the continuous nature of the data, a regression analysis is used to predict the Bradley-Terry values. RMSE is calculated comparing predicted values to the labeled actual value in the test data set. The difference between the predicted value is taken for each prediction. A summation of all the differences is made. That summation is squared and divided by the number of predictions. Finally, the square root of the results of the prior operations is taken. The RMSE is the standard way to evaluate regression models.

There were four models evaluated. The first model was a baseline model using traditional readability metrics, second was a pretrained BERT model, a PyTorch model that combined Readability, Metrics and BERT as an ensemble method, and a Graph CNN model.

The first model used several readability metrics calculated on each text and used as numeric features in a regression model. This was the baseline model for this paper.The readability metrics used were Flesch-Kincaid, Dale-Chall, Gunning Fog, and Coleman-Liau and Ari. In addition, pure word count and lexical diversity were also used to add features to the data. The readability metrics achieved RMSE on the test data set of 0.66. This places a simple regression model based on classic features in the top two-thirds of models submitted to the competition.

The second model attempted to use NLP to process the text excerpts in the data. The baseline NLP model that was chosen was "BERT." BERT has numerous implementations training on different text corpuses and with many different parameters. For this project, the "bert-base-uncased" was used. This model was trained on "Book Corpus" and Wikipedia entries using a Mask Model. As the name indicates, the model disregards whether a word is capitalized or not. For more information regarding the model refer to [36]. The BERT model achieved results of 0.44 for RMSE. This was a marginal improvement over the baseline readability metrics.

The third model hypothesized that an ensemble of numeric features might achieve significantly different results. For this analysis, we used a multi-model transformer model that included both "bert-based-uncased" and the baseline readability metrics. The ensemble method achieved an RMSE value of 0.30. This value is significantly smaller than the two independent models. It also is smaller than competition winners. Based on the RMSE, this proposed model has the best results. Summary metrics are shown for the models evaluated.

The fourth model used a graph convolutional network, which achieved a RMSE of 0.86 on the test dataset. This model used a 50 / 50 train validation split. Accuracy is measured by censoring nodes during evaluation which is different than the data splitting using in the BERT based models.

Table 1: Model Results

Model Comparison	**RMSE**
Baseline Numeric Features Regressor	0.66
Baseline BERT Model	0.44
Proposed Model	0.30
Graph Convolutional Network	0.86

5 Discussion

An automated reading level classification model can be used to quickly determine reading levels for many text documents. The challenge lies in determining if the new model is more accurate than current standards methods such as Flesch-Kincaid. Metrics used to measure the performance of this model were based on the RMSE of the model versus a target value created for data in a Kaggle competition. The target values were created using thousands of pairwise Bradley-Terry comparisons of text excerpts, rated by teachers. This causes the target values to become more subjective when determining the most accurate model.

It is difficult to compare the Flesch-Kincaid method to scores from the Bradley-Terry method, because the Bradley-Terry method results in a ranking given by a probability outcome whereas the Flesch-Kincaid formula gives an actual grade level. Figure 4 below shows a comparison of the Flesch-Kincaid grade levels as calculated from the Readability Python package and the target scores from the training dataset. For grade levels 1-17, they trend together with the target score reducing as the Flesch-Kincaid grade level increases, but it is not a one-to-one match on an individual scale.

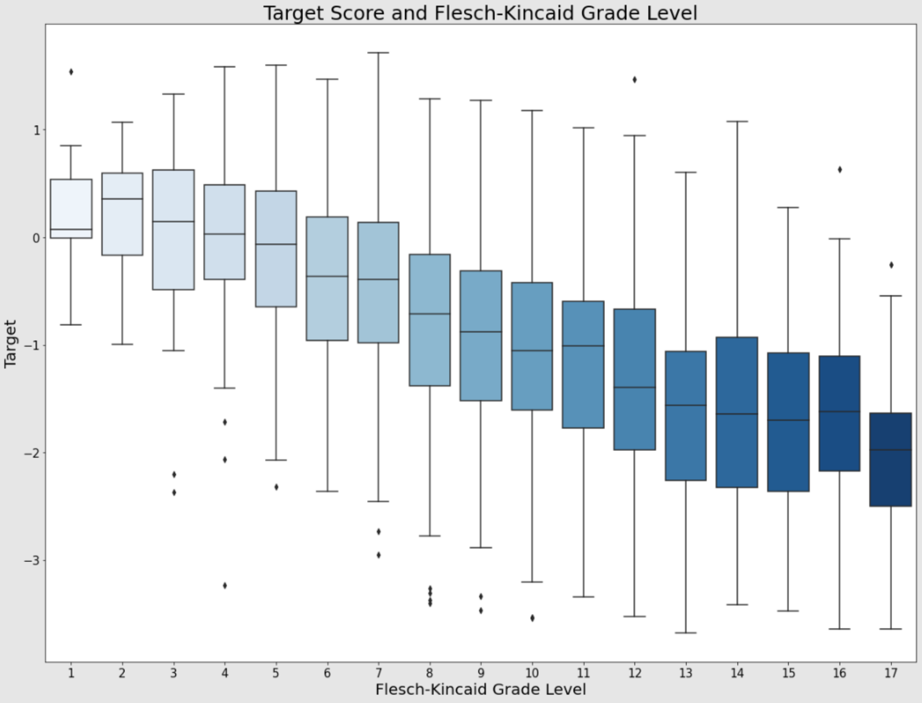

Figure 4: Target Score and Flesch-Kincaid Grade Level. The target score, representing Bradley-Terry Easiness, is negatively correlated with FK Grade Level.

A big advantage of the Flesch-Kincaid model is its simplicity. It is easily explained using simple arithmetic to anyone who wants to understand why a piece of text was classified into a particular grade level. By measuring readability of a document based on a mathematical formula, there is no risk of subjectivity.

In the Kaggle competition, the highest scoring models were complex ensemble models which were extremely computing resource intensive. This could be impractical if we are to create a model to be used by educators on a regular basis. The proposed model that combines NLP and numeric processing was able to get results slightly higher than the highest scoring models, with much less complexity.

Since this problem was presented in a competition, there were many approaches to this task. However, there were not a lot of varied approaches. Review of the competition reveals that most submissions took an approach to train ten to fifty separate BERT models and ensemble them to make a prediction. This resulted in a decrease in the RMSE scores from the baseline BERT score of 0.44 to the 0.41 range.

There are several reasons that this method of solving the problem does not seem to meet the spirit of the competition. The idea behind the reason for the competition is to produce a solution that is usable by educators. A solution that requires the implementor of the resulting production model to train and maintain fifty models is not only wasteful from a computing perspective, but also from a human time investment required to maintain so many models. The solution should produce accurate results, but also do so simply and efficiently. The small amount of performance improvement achieved by ensembles is not significantly different from a simple BERT model and only slightly more effective than classical calculations. The difference between baseline mathematical scores and BERT based models may be as little as six percent.

One area that machine learning can contribute to selecting texts for students is content moderation. There can be texts that are easily read by younger students, but still not appropriate for the age of the reader. A future addition to readability tools should include detection of inappropriate topics including illegal topics or language containing toxic, impolite, suggestive, violent, disturbing, or hateful content also should be excluded from material that is used for educational purposes. There are many current models that support this analysis of language and should be part of any tool provided because of this effort.

The graph model did not achieve results comparable to the other models. The graph has the potential to offer more explainability using the links between similar documents. It is possible this model could be improved using additional hyper parameter tuning, architecture changes, and feature engineering.

6 Ethics

Automated readability measurements do not have a way of eliminating the ethnic or cultural bias of text. This can be a concern because reading levels can be used to evaluate student performance. If a text is determined to be at a certain grade level, students in that grade may be expected to comprehend it. Machine learning algorithms are trained on available text data and cannot consider cultural differences in reading material [42]. Text that is easy to understand for one group of people may not be as easy to understand for another group. A study of American and Iranian students determined that cultural origin of stories was a stronger factor in reading understanding than sentence structure and complexity [42]. If text passages written by authors from minority cultures are given a higher reading level than necessary, the author's voices may not be heard by younger students.

BERT models are known to have gender and ethnic bias. These models tend to associate words related to certain jobs and emotional intensities differently among genders [38]. In addition, words related to careers and activities tend to be associated differently across different ethnic groups [39]. There have been studies to try and identify ways to mitigate this bias, but this was out of scope for reading level identification [38,39]. While BERT bias may not affect the reading difficulty levels directly, it is important to understand that this bias is known to exist. Future research could identify if higher reading levels were associated with text related to certain ethnic and gender groups and investigate if this is problematic.

Automated readability models will not take impolite language into account. It is possible that language within a text passage that may be acceptable for older students would not be acceptable for younger students. If a model determines a reading level for younger readers, there still could be simple language within the passage not appropriate for younger readers

Similar to the issue of impolite language, there could be topics not appropriate for a chosen grade level. The text may be simple to read, and classified for a younger grade level, but it may be related to topics to which young children are not normally exposed.

Because the training data for the model involved human intervention, there could be additional unpredictable bias created by the selection of the teachers who participated in the Bradley-Terry ranking process. As time goes on, the English language evolves. Eventually, the training data could become stale. New participants may be needed.

7 Conclusion

The combination of transformers, readability scores, and linguistic measures are effective at measuring readability as defined by Crossley et al. [1]. Transformers leverage the most current NLP techniques, while readability metrics add information using different statistical measures, which combine to create a more accurate estimate of readability. Flesch-Kincaid scores still have the advantage of simplicity as these new methods come at the cost of significantly increased complexity. This method offers an additional metric of readability which contains more information than simpler methods.

Potential next steps would be to create a front-end interface for the model that would make its use easier for teachers and students. If continued work on the graph model could bring the RMSE close to the proposed model, it may be preferred since it would be a more explainable model with the ability to visualize the edges of the graph. The models would potentially benefit from the addition of linguistic features derived from Wordnet or those suggested by other authors. Finally, the application of explainable methods to these models would help identify the specific textual elements of a document that contribute to its readability score.

Acknowledgments. Jacquelyn Cheun, PhD. – Capstone Professor

References

1. Scott Crossley. (May 2021). CommonLit Readability Prize. 1. Retrieved May 20,2021from https://www.kaggle.com/c/commonlitreadabilityprize/overview.
2. Williamson, J., & Martin, A. (2010). Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. International Journal of Clinical Practice (Esher), 64(13), 1824–1831. https://doi.org/10.1111/j.1742-1241.2010.02408.x
3. Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., & Gabitov, A. (2017). Evaluating Text Complexity and Flesch-Kincaid Grade Level. Journal of Social Studies Education Research, 8(3), 238–248.
4. Derar Eleyan, Abed Othman, & Amna Eleyan. (2020). Enhancing Software Comments Readability Using Flesch Reading Ease Scorexsd. Information (Basel), 11(430), 430–. https://doi.org/10.3390/info11090430
5. The Flesch Reading Ease and Flesch-Kincaid Grade Level (2017) The Readable Blog, https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/
6. Petersen, Sarah E, and Mari Ostendorf. "A Machine Learning Approach to Reading Level Assessment." Computer speech & language 23.1 (2009): 89–106. Web
7. Stockmeyer, N. (2009). Using Microsoft Word's readability program. Michigan Bar Journal, 88(1), 46–.
8. Hoke, Brenda Lynn. "Comparison of Recreational Reading Books Levels Using the Fry Readability Graph and the Flesch-Kincaid Grade Level." N.p., 1999. Print.
9. Fitzgerald, Jill et al. "Important Text Characteristics for Early-Grades Text Complexity." Journal of educational psychology 107.1 (2015): 4–29. Web.
10. Aroyehun, Angel, Alvarez, Alejandro. (2018). Complex Word Identification: Convolutional Neural Network vs. Feature Engineering. Www.Aclweb.Org. https://www.aclweb.org/anthology/W18-0538.pdf
11. Grabeel, Kelsey Leonard et al. "Computerized Versus Hand-Scored Health Literacy Tools: a Comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in Printed Patient Education Materials." Journal of the Medical Library Association 106.1 (2018): 38–45. Web.
12. Maddela, Mounica, and Wei Xu. "A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification." (2018): n. pag. Print.
13. Flesch–Kincaid Readability Tests. (Wikipedia). Available online: https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests (accessed on 18 September 2016).
14. Gong, J., Abhisek, V., & Li, B. (2018). Examining the Impact of Keyword Ambiguity on Search Advertising Performance: A Topic Model Approach. MIS Quarterly, 42(3), 805–829. https://doi.org/10.25300/MISQ/2018/14042
15. Ojha, Pawan Kumar, Abid Ismail, and K.S Kuppusamy. "Perusal of Readability with Focus on Web Content Understandability." Journal of King

Saud University. Computer and information sciences 33.1 (2018): 1–10. Web.

16. Matej Martinc, Senja Pollak, and Marko Robnik-Sikonja. "Supervised and unsupervised neural approaches to text readability." arXiv Cornell University(2021). https://arxiv.org/abs/2104.13103
17. Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. Journal of Research in Reading, 42 (3-4), 541-561. https://alsl.gsu.edu/files/2019/08/moving-beyond-classic-readability-formulas-off-print.docx
18. Vaswani, Ashish et al. "Attention Is All You Need." (2017): n. pag. Print. https://arxiv.org/pdf/1706.03762.pdf
19. Horev, Rani "BERT Explained: State of the art language model for NLP"(2018)//towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
20. Dale, E., & Chall, J. (1949). The Concept of Readability. Elementary English, 26(1), 19-26. Retrieved June 25, 2021, from http://www.jstor.org/stable/41383594
21. Kemper, S. (1983). Measuring the inference load of a text. Journal of Educational Psychology, 75(3), 391–401. https://doi.org/10.1037/0022-0663.75.3.391
22. Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329
23. JS Chall & E. Dale (1995) Readability revisited: The new Dale-Chall readability formula
24. Bonnie J. F. Meyer. (1982). Reading Research and the Composition Teacher: The Importance of Plans. College Composition and Communication, 33(1), 37-49. doi:10.2307/357843
25. Amalia Todirascu, Thomas François, Delphine Bernhard, Núria Gala, Anne-Laure Ligozat. Are Cohesive Features Relevant for Text Readability Evaluation?. 26th International Conference on Computational Linguistics (COLING 2016), Dec 2016, Osaka, Japan. pp.987 - 997. ⟨hal-01430554⟩
26. "Developmental Reading Assessment: Third Edition." Pearson, www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Academic-Learning/Developmental-Reading-Assessment-|-Third-Edition/p/100001913.html?tab=overview.
27. Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. The Elementary School Journal, 115(2), 184–209. https://doi.org/10.1086/678294
28. Amalia Todirascu, Thomas François, Delphine Bernhard, Núria Gala, Anne-Laure Ligozat. Are Cohesive Features Relevant for Text Readability Evaluation?. 26th International Conference on Computational Linguistics (COLING 2016), Dec 2016, Osaka, Japan. pp.987 - 997. ⟨hal-01430554⟩
29. U.S. Department of Education https://www2.ed.gov/policy/elsec/leg/essa/index.html

30. Bradley, Ralph Allan and Terry, Milton E. "Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons" Biometrika 39.3-4 (1952): 324–345. Web.
31. National Association of Educational Progress (2019) NAEP Report Card: 2019 NAEP Reading Assessment, Highlighted results at grades 4 and 8 for the nation, states, and districts https://www.nationsreportcard.gov/highlights/reading/2019/
32. National Association of Educational Progress (2019) NAEP Report Card: 2019 NAEP Reading Assessment, Highlighted results at grade 12 for the nation. https://www.nationsreportcard.gov/highlights/reading/2019/g12/
33. DiMascio, Carmine "Determine the "Readability" of a text with Python" (2019) https://levelup.gitconnected.com/determine-the-reading-level-of-a-text-with-python-d2f9dccee6bf
34. "Readability Formulas"(2021) Web. https://readabilityformulas.com/
35. Michael B.W. Wolfe, M.E. Schreiner, Bob Rehder, Darrell Laham, Peter W. Foltz, Walter Kintsch & Thomas K Landauer (1998) Learning from text: Matching readers and texts by latent semantic analysis, Discourse Processes, 25:2-3, 309-336, DOI: 10.1080/01638539809545030
36. Bert-base-uncased · hugging face. bert-base-uncased · Hugging Face. (n.d.). Retrieved October 3, 2021, from https://huggingface.co/bert-base-uncased.
37. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*..
38. Bhardwaj, Rishabh, Navonil Majumder, and Soujanya Poria. "Investigating Gender Bias in BERT." Cognitive computation 13.4 (2021): n. pag. Web.
39. Ahn, Jaimeen, and Alice Oh. "Mitigating Language-Dependent Ethnic Bias in BERT." (2021): n. pag. Print.
40. Lu, Z., Du, P., & Nie, J.-Y. (2020). VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In Advances in Information Retrieval (pp. 369–382). Springer International Publishing. https://doi.org/10.1007/978-3-030-45439-5_25
41. Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks.
42. Johnson, Patricia. "Effects on Reading Comprehension of Language Complexity and Cultural Background of a Text." TESOL quarterly 15.2 (1981): 169–181. Web.
43. Mridha, Sankarshan. "Survey - BERT". (2020) https://msank00.github.io/blog/2020/04/13/blog_607_Survey_BERT