Report: A Feature-Based Approach to Predicting Text Readability

1. Problem Definition

This project addresses the Natural Language Processing problem of text readability prediction. This is a regression task aiming to score the ease of comprehension of a text, often mapped to a continuous score. Accurate readability prediction is crucial for developing educational materials, ensuring content accessibility, and matching content to appropriate audiences.

The project utilizes the CommonLit Readability Dataset, a "real" dataset of 2,834 excerpts from elementary to college-level educational texts.

Datasource: [CommonLit Readability Prize] (https://www.kaggle.com/competitions/commonlitreadabilityprize/data)

2. Methodology and Model

Our solution consists of a feature-driven machine learning pipeline using a Random Forest Regressor as the core model.

The novelty of this approach lies in moving beyond classic formulas (which rely only on surface statistics) by integrating semantic context (via bigram embeddings) with traditional statistical and lexical (TF-IDF) features.

We constructed our feature set in three phases:

(1) Feature Set 1 (Baseline): Statistical Features

- Captures foundational text structure, e.g., average word/sentence length, lexical diversity, and classic `textstat` metrics (Flesch-Kincaid, etc.).

(2) Feature Set 2: Lexical Features (TF-IDF)

- Uses `TfidfVectorizer` (with `max_features=1000` and `ngram_range=(1, 2)`) to identify discriminative words and phrases correlated with text complexity.

(3) Feature Set 3: Semantic Embeddings (Bigram W2V)

- Employs `gensim.Phrases` to identify common bigrams (e.g., "high_school") from the training text.
- Uses a pre-trained `glove-wiki-gigaword-100` word vector model to extract 100-dimensional vectors for all words and phrases.

- Generates a single vector per document by taking the mean of all constituent vectors, representing its overall semantic meaning.

Note on Model Type: Our Random Forest Regressor is a discriminative model that learns the direct mapping from features to readability scores. As it does not make probabilistic assumptions about data generation, Step 3a (evaluation on synthetic data) does not apply to this project. We proceed directly to Step 3b with real-world data evaluation.

3. Evaluation

We used Root Mean Squared Error (RMSE) as the primary quantitative evaluation metric.

3.1. Quantitative Evaluation

We trained models by progressively adding feature sets to demonstrate the contribution of each component. All results are reported on a 20% held-out test set.

Core Project Results Table

Model Phase	Features Used	RMSE (Lower is Better)
Phase 1 (Baseline)	Statistical Only	0.8884
Phase 2	Statistical + TF-IDF	0.7891
Phase 3 (Final Model)	**Statistical + TF-IDF + Semantic**	**0.6997**

Conclusion: The data clearly shows that adding TF-IDF lexical features provided a significant 11.2% reduction in RMSE. Most importantly, adding our novel semantic embedding feature (Phase 3) reduced the error by another 11.3%, validating our core hypothesis.

3.2. Qualitative Evaluation

We performed an analysis of the examples where the model performed best and worst:

(1) Examples Where the Model Performed Well

- The model was nearly perfect (`abs_error < 0.01`) at predicting standard informational text (e.g., an encyclopedic entry on "Seismology").
  - Excerpt Sample:

  Seismology is the study of what is under the surface of the Earth by measuring vibrations on the Earth's surface. A person who does this is called a seismologist. It is part of the science of geophysics, which studies the physics of the processes that formed the Earth and other planets. Seismology is done by seismologists and geophysicists using devices to pick up the vibrations called geophones, hydrophones or seismometers.

Seismology can either be passive, just listening to vibrations caused by earthquakes and volcanic activity, or active, using small explosive charges to send vibrations into the ground. Seismic detectors come in two types, one which measures up and down vibrations, and one which measures side to side vibrations.

Both types use and arrangement of a magnet and a coil of wire which will convert the vibrations into an electrical signal which can be stored in a computer for analysis. Seismologists can find the location of earthquakes by plotting received vibrations on a map. They can also pick up underground nuclear tests, and this is what many of the seismic recording stations were set up for.

target	predicted_target	error
-1.154857	-1.151876	**0.002981**

- **Reason:** This type of text is highly common in the training set. The model is in its "comfort zone" where the statistical, lexical, and semantic features are all in agreement.

(2) Examples Where the Model Performed Poorly

- **Failure Mode 1 (Failure to Extrapolate):** The model systematically fails on texts with extreme difficulty (e.g., `target = -3.23`). It predicts a moderate difficulty (e.g., `-1.22`) because it lacks sufficient extreme examples in the training data to learn how to extrapolate to this range.
    - Excerpt Sample:

A first-class boat will be of about the following dimensions: Length over all, 45 ft. to 50 ft.; breadth (extreme), 9 ft. to 10 ft. 3 in.; depth (inside), 3ft. 10 in. to 4 ft. The keel is of oak 6 in. by 3½ in. The stem and stern posts are also of oak. The planking is generally of oak or walnut —the latter preferred—and is 3 in. thick, the width of the planks being 4½ in. Many boats are now constructed of hard wood to the water line and Norway pine above.

The fastenings are galvanized nails 4½ in. long. The mast-partners and all the thwarts are of oak 1½ in. thick and 8 in. wide; the latter are fastened in with iron knees. Lee-board and rudder are of oak, walnut, or chestnut; the rudder extends 3½ ft. to 4 ft. below the keel, and, in giving lateral resistance, balances the lee-board, which is thrust down forward under the lee-bow. The rig consists of two lags, the smaller one forward right in the eyes of the boat; the mainmast being amidships.

target	predicted_target	error
-3.236543	-1.221065	**2.015478**

- **Failure Mode 2 (Feature Confusion):** The model fails on simple biographies with many proper nouns (e.g., "Helen Adams Keller...", `target = +0.64`), incorrectly predicting it as "difficult" (`-1.12`).
  - Excerpt Sample:

  Helen Adams Keller was born on 27th June 1880, in Tuscumbia, Alabama, United States. Her family lived on an estate called Ivy Greens, built by Helen's grandfather. Her father, Arthur Keller, spent many years as an editor for the Tuscumbia North Alabamian newspaper, and had served as a captain in the Army. Her mother, Kate Adams, was the daughter of a confederate general. Helen was born with the ability to see and hear. At 19 months old, she became ill, and this illness left Helen both deaf and blind.

  As she grew up, she found a way of communicating with the daughter of the family's cook; Martha Washington. They invented a kind of sign language and by the time Helen was 7 years old they had created more than 60 different signs for use in their personal communication.

  Around this time, Helen became very frustrated and diffcult to control. She had violent temper tantrums and would giggle uncontrollably when she was happy. Her family was worried about Helen and went in search of help. Unaware of how to deal with Helen's disabilities, the family had indulged , which at this point it was to her detriment.

target	predicted_target	error
0.646549	-1.126084	**-1.126084**

- **Reason:** The TF-IDF features flagged "Keller" and "Adams" as "rare words," and the model has incorrectly learned to associate "rare words" with "high difficulty."

4. Discussion

Based on the evaluation, we summarize the model's pros and cons:

(1) Pros

- **Correctness:** The final model (RMSE: 0.6997) is highly accurate, and we demonstrated the value of our feature combination quantitatively.

(2) Cons

- **Requirements:** The Phase 3 model is computationally heavier, requiring the loading of a large (~130MB) pre-trained GloVe model and a slow batch process to generate the mean vectors.
- **Interpretability:** Interpretability decreases as complexity increases. The 100-dimensional semantic vectors (Phase 3) are a "black box," making it hard to debug cases like the "Helen Keller" failure.

- **Limitations:** As shown in qualitative analysis, the model performs poorly on out-of-distribution (extreme) values and is susceptible to being misled by proper nouns.

5. References

CommonLit. (2021). CommonLit readability prize dataset. Kaggle. https://www.kaggle.com/competitions/commonlitreadabilityprize/data

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). ELRA. http://is.muni.cz/publication/884893/en

Shoemaker, K. (2014). textstat: Python package for calculating statistics from text to determine readability, complexity and grade level [Computer software]. GitHub. https://github.com/textstat/textstat