Language Agnostic Readability Assessments (LARA)

Mrinmoy Dalal, Sankarsan Gautam, Vedanti Gulalkari, Shreyas Joshi, Venkatesh Seetha, Amal Tom, Matthew A. Lanham

Purdue University, Daniels School of Business

dalalm@purdue.edu; gautam15@purdue.edu; vgulalka@purdue.edu; joshi211@purdue.edu; vseetha@purdue.edu; tom3@purdue.edu; lanhamm@purdue.edu

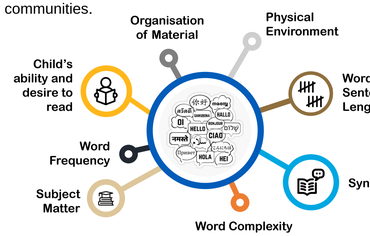
SIL

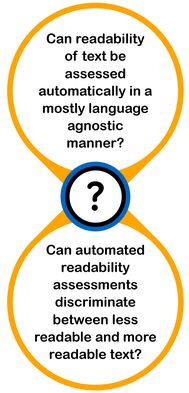
PURDUE UNIVERSITY
Mitchell E. Daniels, Jr. School of Business

informs BUSINESS ANALYTICS CONFERENCE

BUSINESS PROBLEM

Effective communication depends on readability - the ease with which a reader can understand a message. Unfortunately, most readability assessment frameworks are only available for specific languages, which presents a significant challenge for organizations seeking to promote literacy and communication in local language communities.



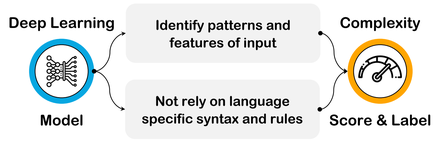


To address this issue, SIL - a global non-profit organization specializing in language development and literacy – is seeking a python-based, language-agnostic readability assessment model that can distinguish simple sentences from complex ones and assign a readability score, regardless of specific language syntaxes or semantics.

This solution can benefit educators, language communities, and individuals seeking to improve their literacy skills by offering a more inclusive tool to assess the readability of written content. It will enable greater engagement across language and literacy levels, facilitate targeted dissemination of knowledge, and help development of digital resources for local language communities.

ANALYTICS PROBLEM

Traditional readability metrics are limited to specific languages and rely heavily on sentence and word lengths, making them less effective for assessing readability across different languages and cultures. Assuming that there are certain patterns and features that are common to all languages, such as the use of subject-verb-object structures, it may be possible to engineer a machine learning model that could capture the unique features of each language while also identifying the common factors that contribute to readability.

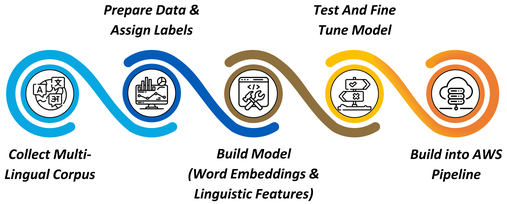


Success Metrics

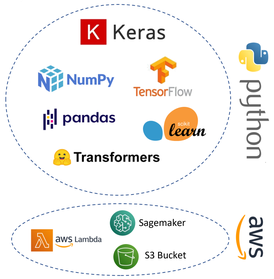
- Model accuracy
- Averaged F1 scores
- Recall
- Precision

Across both target classes by language

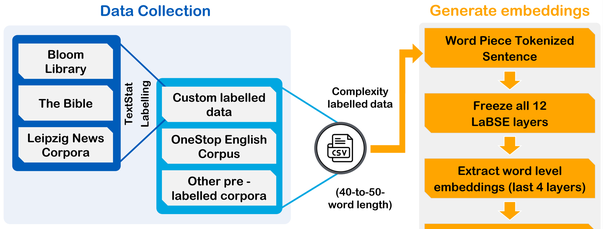
METHODOLOGY



Our methodology leverages language-agnostic sentence embeddings from a large language model to facilitate transfer learning using lexical and semantic information. We train supervised deep-learning models on labeled sentences from major language families and employ zero-shot learning to predict the readability class of multi-lingual verses. By conducting training and prediction in the embedding space, the need for costly language-specific processing is eliminated.



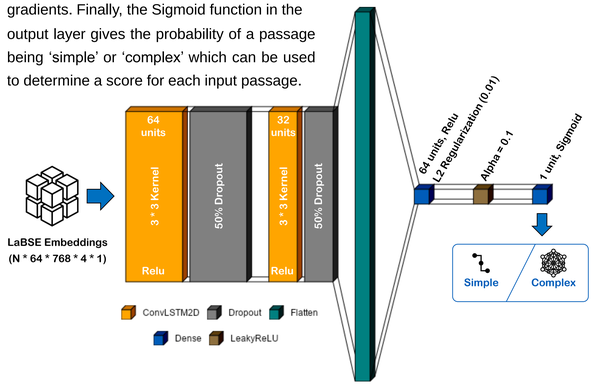
DATA COLLECTION AND PROCESSING

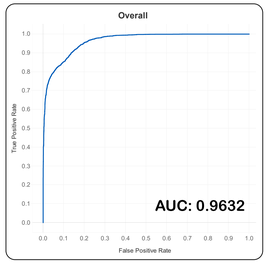


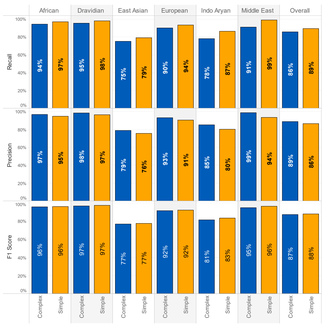
We leveraged pre-labelled data used in other research publications and annotated additional data using the TextStat python package. The input is limited to 40–50 words to simulate the average length of a paragraph of text. This dataset provides a diverse range of sentence structures for our model to learn from without picking up on the length as the obvious differentiator between simple and complex. This data was then pre-processed and converted into LaBSE embeddings which is used to train a deep neural network model for further classification task.

MODEL BUILDING AND RESULTS

The output from last 4 layers of LaBSE are fed into ConvLSTM2D layers to analyze the spatio-temporal features of the word embeddings. This allows for a deep understanding of the context in which the words are used, improving the accuracy of downstream tasks. To prevent overfitting to the training data, L2 regularization is employed during the training of the deep neural network. The activations of the dense layers are then processed by a Leaky ReLU unit, which helps to alleviate the problem of vanishing gradients. Finally, the Sigmoid function in the output layer gives the probability of a passage being 'simple' or 'complex' which can be used to determine a score for each input passage.

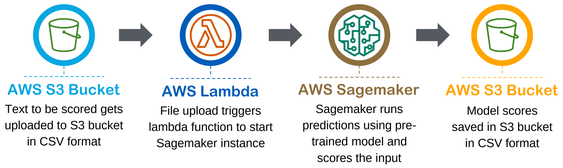






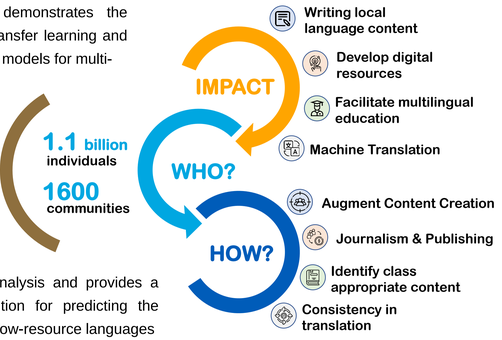
The ROC curve above has an AUC score of 96% and quantifies the model's performance on classifying the complexity of passages. The model performs exceedingly well for African, European, Middle-Eastern and Dravidian languages. For East Asian languages the performance is average at F1-score of 77%, primarily due to the sub-par performance on Japanese text. Fine-tuning the LaBSE model on the specific classification task can improve the quality of the embeddings but requires additional data and is currently out of scope.

DEPLOYMENT & LIFECYCLE MANAGEMENT



To deploy the solution at scale, we developed an AWS pipeline that automatically generates predictions based on new CSV files uploaded to an S3 bucket. The pipeline is designed to be scalable, secure, and cost-effective, and leverages SageMaker notebook instances and lifecycle configurations to automate the setup and configuration of the required dependencies and libraries.

Our project demonstrates the potential of transfer learning and deep learning models for multi-lingual text analysis and provides a practical solution for predicting the readability of low-resource languages using a language-agnostic approach. We hope that our work inspires further research and innovation in this area and contributes to the development of more accessible and inclusive language technologies.



ACKNOWLEDGEMENTS

We would like to thank our industry partner, SIL, for their guidance and support on this project as well as the Purdue MS BAIM program for partially funding this work.