Language Agnostic Readability Assessments

Mrinmoy Dalal
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
dalalm@purdue.edu

Sankarsan Gautam
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
gautam15@purdue.edu

Vedanti Gulalkari
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
vgulalka@purdue.edu

Shreyas Joshi
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
joshi211@purdue.edu

Venkatesh Seetha
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
vseetha@purdue.edu

Amal Tom
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
tom3@purdue.edu

Matthew A. Lanham
Mitchell E. Daniels Jr. School of Business, Purdue University
West Lafayette, USA
lanhamm@purdue.edu

***Abstract* — Readability assessment models are increasingly important in evaluating the complexity of written text and ensuring it is suitable for a specific audience. Automatic readability assessment models are particularly useful for educators, content creators, and researchers who want to make their materials easily understandable and accessible. This paper presents a novel approach for language-agnostic automatic readability assessment using a combination of machine learning techniques and natural language processing tools. The proposed method was tested on a diverse dataset of texts in multiple languages and demonstrated strong performance in accurately assessing readability. This approach provides a valuable tool for working with multilingual text and can help bridge the gap in readability assessment for languages that lack dedicated tools.**

Keywords - NLP, readability assessment, language-agnostic, LaBSE, unstructured data

I. INTRODUCTION

Readability assessment models provide a way to measure the complexity of text and determine its suitability for a particular audience, such as students, the public, or experts in a field. By providing a systematic method for evaluating the readability of text, these models help to ensure that written material is accessible and easily understandable for all readers. Globalization has brought about a significant increase in the need for clear and concise communication across linguistic borders. With a diverse and multinational workforce, businesses, and organizations require materials that are communicable to a wide range of individuals speaking diverse languages. This has created a need for readability assessments that are language-agnostic. These assessments ought to improve on existing metrics by measuring the ease of understanding a text, not just in terms of grammar and vocabulary but also in terms of its structure and style, making them suitable for use across different languages and cultures. With a language-agnostic readability assessment, organizations can ensure that their communications are clear, effective, and accessible to all.

The proposed approach for automatic language-agnostic readability assessment of text is based on a combination of machine learning techniques and natural language processing (NLP) tools, which allows for an accurate and precise evaluation of the readability level of written material. To assess its effectiveness, we tested the method on a vast and diverse dataset of texts in 29 different languages, spanning across various language families. The obtained results demonstrate a strong performance in accurately assessing readability, providing a valuable and practical tool for educators, content creators, and researchers working with multilingual text. Notably, the approach is highly flexible and language agnostic, enabling its application to any language and addressing the gap in readability assessment tools for languages that do not have dedicated resources.

One example of the use of this language agnostic readability assessment model is in the field of journalism. Many renowned newspapers, such as The New York Times, The Guardian, and Le Monde, use readability assessment models to evaluate the readability of their articles before publishing them. However, these newspapers have articles in multiple languages. For example, The Guardian uses readability assessment models to measure the complexity of their articles to ensure that they are written for a reading age of 15. The multilingual applicability of this approach ensures that this requirement is met regardless of the language of the article.

The limitations of existing readability metrics, which have been developed and refined by educators and researchers since the 20th century, are becoming increasingly apparent. These metrics, such as the Flesch-Kincaid and Gunning Fog Index, rely on simple metrics like word and sentence length, along with parts-of-speech (POS) specifics like prepositions and adverbs, to evaluate the complexity of

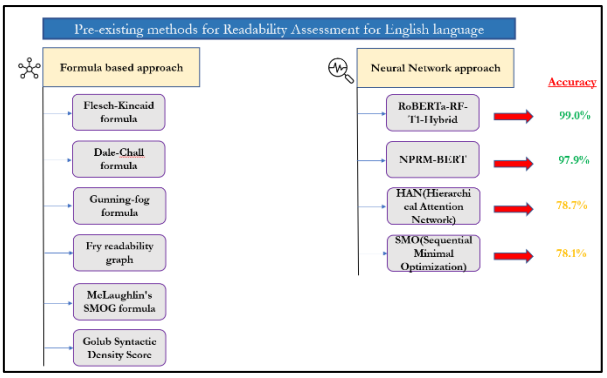

Figure 1 – Existing readability assessment methods for English language

written material. However, they suffer from several drawbacks, including their language-specific nature, overemphasis on length-based metrics, lack of agreement on what constitutes a difficult word, inability to handle scripts without whitespace separation, domain specificity, and failure to consider the context and dynamic nature of languages. These limitations necessitate a new approach that can address these issues, overcome their restricted applicability and accuracy, and provide a more comprehensive and flexible tool for assessing the readability of written material.

With the advancement of natural language processing (NLP) techniques, we are now able to better assess the readability of text by overcoming these challenges. NLP-based readability assessment models can analyze text at a deeper level, capturing complex non-linear features, while factoring in aspects such as grammar, vocabulary, and semantics. Additionally, the use of machine learning techniques allows these models to be trained on increasingly large amounts of data, keeping up with the evolution of languages, increasing their accuracy and making them more robust.

The figure below displays the major languages in the world by number of people speaking them. The languages shown by orange bar are the ones that have their tokenizers and POS taggers and thus can be easily fed to a ML model. These encompass about 45% of people in the world. With the help of NLP, we can take readability assessment to the next level, providing a more accurate, efficient, and **language-agnostic** method for evaluating text.

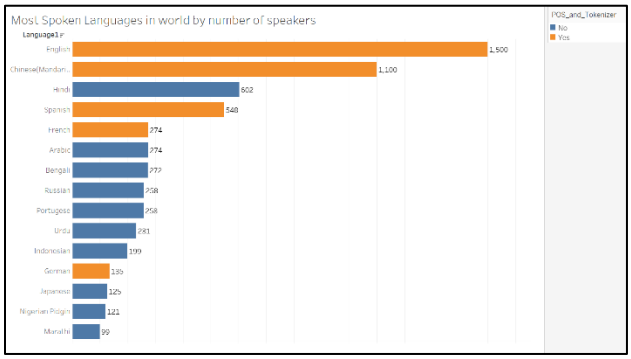

Figure 2 – Most spoken languages in the world by number of speakers

The remainder of this paper is organized as follows:

Section 2: we provide a description of the data used to train our model and evaluate its accuracy and performance on various languages.
Section 3: we present our methodology for implementing the solution for different sets of languages, considering the features shared across different languages
Section 4: explains the usage of LaBSE embeddings to train a deep neural network model to capture the spatio-temporal nature of the input text.
Section 5: we present the performance of our models based on various metrics.
Section 6: we conclude the paper with a discussion on the implications of our study, potential future research directions, and concluding remarks.

II. DATA

We leveraged pre-labelled data used in other research publications such as the OneStopEnglish corpus and annotated additional data from the Leipzig Corpora news dataset 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 to force the model to learn features beyond the length as the dominant classification factor.

In this study, we used the language data provided by the client to test our model. The data consists of iso code of languages, the type of script, whether their words are separated by whitespace, a complicated sentence in that language, and a simple sentence. Table 1 below provides a brief description of the test dataset used in this study.

Data Dictionary	
iso	ISO Language code for the two sentences
script_type	type of script, whether it contains alphabets, symbols etc. (alphabet, abjad, abugida, featural, logo-syllabary)
white_space	delimiters between words (yes, no)
complicated	Complicated version of the sentence
simple	Simple version of the sentence

Table 1 – Test data dictionary

III. METHODOLOGY

Given the limited availability of language-specific text pre-processing resources, such as tokenizers and POS taggers, for low-resource languages, our methodology follows the feature-based approach for using transformer outputs. Our setup relies on language-agnostic sentence embeddings from a large language model to provide features capturing rich lexical and semantic information, thereby facilitating effective transfer learning.

In order to enhance the zero-shot learning performance, it is vital to train the downstream DNN on feature maps from multiple languages. This approach enables the model to effectively distinguish the nuances and similarities among languages and develop a generalized logic, which can then be applied to distinguish between simple and complex passages in unseen yet related test languages. To achieve this supervised learning approach, the raw datasets were

labeled using traditional readability metrics during the training process. These datasets were enriched with pre-labeled open-source data available for different languages, with the aim of increasing the diversity of training samples. It is worth noting that text pre-processing was kept minimal, as language models are typically trained on unprocessed text and trade-off higher dimensionality for a better understanding of context.

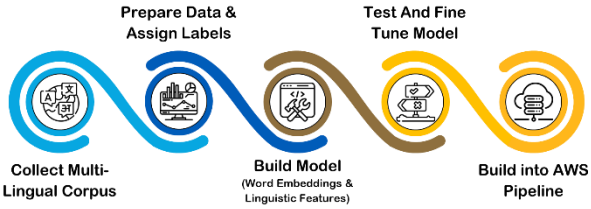

Figure 3 – Methodology

The design of the deep neural network architecture underwent several iterations before the current version was settled upon. Our primary goal was to extract increasingly nuanced information from the LaBSE embeddings, achieved by implementing deep filters to enable the dense layer to learn from selectively curated features. Our exploration included a Multi-channel CNN architecture with a focus on identifying n-gram features, as well as fine-tuning LaBSE to optimize embedding quality. However, after carefully considering model stability and production constraints, the final approach discussed below produced the most favorable results.

Given that this is essentially a binary classification problem, the evaluation metrics employed are similar to those used for other problems in this domain, including accuracy, precision, recall, F1, and AUC scores. To evaluate the model, we perform an 80-20 train-test split of the input dataset. Furthermore, we monitor the validation loss after each epoch of training to prevent overfitting and determine the optimal training cycles. To validate the performance of zero-shot learning, the test set contains passages from both languages that the model has been trained on and those it has not.

Our approach to deploying the solution is to create a seamless pipeline in AWS. The pipeline is comprised of several components, each serving a unique purpose:

1) **S3 Bucket:** The CSV file containing passages to be scored is uploaded into an S3 bucket: *purdue-ip-2023-readability/sil2-pred/*
2) **Lambda Function:** AWS lambda function (*Sil_readability_v1*) is triggered when a new CSV file is uploaded to the S3 bucket. The function starts a SageMaker notebook instance.
3) **SageMaker Notebook Instance:** A SageMaker notebook instance (*sil2-ram-optimised*) is started by the lambda function and is configured with a lifecycle configuration.
4) **SageMaker Lifecycle Configuration:** A set of scripts that automate the setup and configuration of the SageMaker notebook instance. The lifecycle configuration (*lifecycle-test-3*) sets up the required dependencies and libraries required to run a particular notebook.
5) **Prediction Notebook:** A Jupyter notebook that generates predictions based on the input CSV file. The notebook *(lambda_pred.ipynb*) is executed automatically by the lifecycle configuration when the SageMaker notebook instance starts.

Our pipeline is designed to be scalable, secure, and cost-effective. It is triggered automatically when a new CSV file is uploaded to the S3 bucket and produces predictions quickly and efficiently using SageMaker. Furthermore, the pipeline automatically stops the SageMaker Notebook Instance after 10 minutes of inactivity.

IV. MODELS

The model being deployed is a supervised learning algorithm. The features for the modelling task are being extracted as embeddings for each input passage as generated using the LaBSE Large Language Model (LLM). These embeddings are in the form of a 3D matrix for each passage in the training dataset. Each such matrix is assigned either a ‘simple’ or ‘complex’ label, determined using expert judgement. Subsequently this embedding matrix and corresponding label are used to train a deep neural network on a downstream classification task which involves distinguishing between ‘simple’ and ‘complex’ passages. Over enough training cycles, the deep neural network is able to capture the nuanced difference in embeddings of different passages that determine whether they are ‘simple’ or ‘complex’.

The deep neural network configuration consists of a layer of 64 ConvLSTM2D filters, followed by a 50% dropout layer, which is then followed by a layer of 32 ConvLSTM2D filters and another 50% Dropout layer, with layers of filters having ReLU activation functions. The output of these layers is then flattened before being passed to a fully connected dense layer with 64 hidden units and L2 regularization. The activations of the dense layers are ultimately passed to an output layer with binary cross-entropy loss function, going through a Leaky ReLU unit. The model is trained over 10 epochs with a batch size of 64. 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.

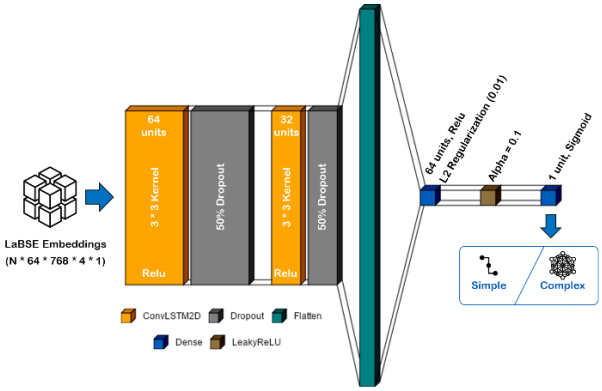

Figure 4 – Model Architecture

The Dropout rate of 50% and the L2 regularization together are meant to tackle overfitting on the training dataset. To further mitigate this issue, the training process was limited to 10 epochs. Two primary factors that can contribute to overfitting are the model's complexity and the limited size of the input dataset. A model with a sizeable parameter set will memorize the training data over multiple iterations, leading to high training accuracy but low accuracy on test inputs. The model also suffered from vanishing gradients, which necessitated the Leaky ReLU unit after dense layers. Ultimately, having taken measures against overfitting and vanishing gradients, we were able to improve the precision, recall and F1 scores on the test passages.

The LaBSE embeddings have been utilized as input features for training the model in this case. However, it is possible to further improve the quality of the embeddings by fine-tuning the base LaBSE model on our particular classification task. This requires additional training data and is out-of-scope for now. There is merit in performing further research on the potential of finetuned embeddings in improving the current model.

V. RESULTS

The model's performance was validated using precision, recall, and F1 scores. The evaluation criteria judged its capability on a variety of languages, specifically the extent to which its able to correctly classify the complexity of multilingual passages.

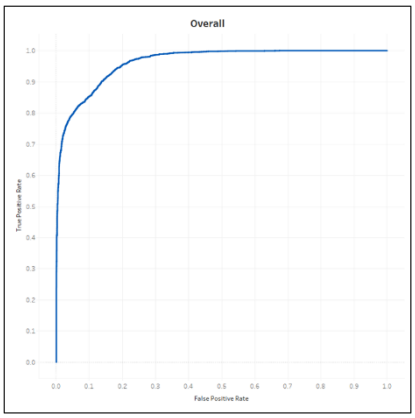

Figure 5 – AUC curve

The ROC curve above has an AUC score of 96% and quantifies the model's performance on classifying the complexity of passages. Even though this demonstrates the model's overall performance, we wanted to evaluate its capability with respect to specific language families. The results below indicate the capabilities of the model for each of the 2 classes, given each of the language families:

For 'Simple' class:

Language Family	Precision	Recall	F1-score
African	95%	97%	96%
East Asian	76%	75%	77%
Indo-Aryan	80%	87%	83%
European	91%	94%	92%
Middle Eastern	94%	99%	96%
Dravidian	97%	95%	97%
Overall	86%	86%	88%

Table 2 – Results for 'simple' sentences

For 'Complex' class:

Language Family	Precision	Recall	F1-score
African	97%	94%	96%
East Asian	79%	75%	77%
Indo-Aryan	85%	78%	81%
European	93%	90%	92%
Middle Eastern	99%	91%	95%
Dravidian	98%	95%	97%
Overall	89%	86%	87%

Table 3 – Results for 'complex' sentences

As observed above, 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. Given the limited labelled data that was collected for Japanese, this can be remedied by using more data for training. Overall, the model is able to satisfactorily capture the distinction between complex and simple sentences as per the defined criteria.

VI. CONCLUSION

In this project, we developed an approach evaluate the readability of multi-lingual texts using a transfer learning and deep learning architecture trained on labeled English Bible, Leipzig Corpora, OneStop English Corpus and Bloom verses. By conducting training and prediction in the embedding space, the need for costly language-specific processing is eliminated, effectively reducing the expenses and complexities associated with developing models for low-resource languages.

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 results show that our model can accurately predict the readability of multi-lingual texts, achieving an overall accuracy of 87%, overall precision of 88%, and overall recall of 85% across 29 test languages. We believe that this approach can be extended to other applications and domains, and that it can help overcome the challenges of developing models for low-resource languages.

In summary, 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.

REFERENCES

1) Yang, Ziyi, et al. *Universal Sentence Representation Learning with Conditional Masked Language Model*. 10 Sept. 2021, https://arxiv.org/pdf/2012.14388.pdf.

2) Feng, Fangxiaoyu, et al. "Language-Agnostic Bert Sentence Embedding." *ArXiv.org*, 8 Mar. 2022, https://arxiv.org/abs/2007.01852.

3) Imperial, Joseph Marvin. "Bert Embeddings for Automatic Readability Assessment." *ArXiv.org*, 30 July 2021, https://arxiv.org/abs/2106.07935.

4) Martinc, Matej, et al. *Supervised and Unsupervised Neural Approaches to Text Readability*. 26 July 2019, https://aclanthology.org/2021.cl-1.6.pdf.

5) Deutsch, Tovly, et al. "Linguistic Features for Readability Assessment." *ACL Anthology*, 20 July 2020, https://aclanthology.org/2020.bea-1.1/.

6) Vajjala, Sowmya. "Trends, Limitations and Open Challenges in Automatic Readability Assessment Research." *ArXiv.org*, 19 Apr. 2022, https://arxiv.org/abs/2105.00973.

7) Ortiz-Zambrano, Jenny A., et al. "Combining Transformer Embeddings with Linguistic Features for Complex Word Identification." *MDPI*, Multidisciplinary Digital Publishing Institute, 27 Dec. 2022, https://www.mdpi.com/2079-9292/12/1/120.

8) Azpiazu, Ion Madrazo, and Maria Soledad Pera. "Multiattentive Recurrent Neural Network Architecture for Multilingual Readability Assessment." *ACL Anthology*, 1 Jan. 1970, https://aclanthology.org/Q19-1028/.

9) Vajjala, Sowmya, and Ivana Lučić. "OneStopEnglish Corpus: A New Corpus for Automatic Readability Assessment and Text Simplification." *ACL Anthology*, https://aclanthology.org/W18-0535/.

10) "Corpora Collection." *Downloads*, https://wortschatz.uni-leipzig.de/en/download.