Under the Microscope: Interpreting Readability Assessment Models for Filipino

Joseph Marvin Imperial
National University
Manila, Philippines
jrimperial@national-u.edu.ph

Ethel Ong
De La Salle University
Manila, Philippines
ethel.ong@dlsu.edu.ph

Abstract

Readability assessment is the process of identifying the level of ease or difficulty of a certain piece of text for its intended audience. Approaches have evolved from the use of arithmetic formulas to more complex pattern-recognizing models trained using machine learning algorithms. While using these approaches provide competitive results, limited work is done on analyzing how linguistic variables affect model inference quantitatively. In this work, we dissect machine learning-based readability assessment models in Filipino by performing global and local model interpretation to understand the contributions of varying linguistic features and discuss its implications in the context of the Filipino language. Results show that using a model trained with top features from global interpretation obtained higher performance than the ones using features selected by Spearman correlation. Likewise, we also empirically observed local feature weight boundaries for discriminating reading difficulty at an extremely fine-grained level and their corresponding effects if values are perturbed.

1 Introduction

Readability assessment is the process of evaluating a certain piece of text or reading material in terms of reading difficulty. Likewise, reading difficulty can be expressed in various forms such as age level, grade level, or a number from a certain range defined by a book publisher (Deutsch et al., 2020). Through the years, this process has evolved from the use of handcrafted arithmetic formulas such as the Flesch-Kincaid Reading Ease (Kincaid et al., 1975) and Dale-Chall (Dale and Chall, 1948) readability formulas to the use of supervised machine learning algorithms such as Logistic Regression and Support Vector Machines (Chatzipanagiotidis et al., 2021; Weiß and Meurers, 2018; Xia et al., 2016; Reynolds, 2016; Vajjala and Meurers, 2012). Despite the significant growth in research history, several problems still pose as open challenges for the task such as the (a) availability of corpora and tools for linguistic feature extraction, (b) extrinsic evaluation, and (c) interpretation of linguistic predictors used which is arguably the most important of all (Vajjala, 2021).

We describe the general readability assessment task as *feature interpretation dependent* since identifying which optimal subset of linguistic features that can potentially impact readability levels is a strict and necessary part of the research process that should not be ignored. Recent works in readability analysis (Imperial and Ong, 2020a; Hancke et al., 2012) have used standalone, correlation-based feature selection techniques such as Spearman or Pearson correlation to get a better understanding of feature dependence and relationship. These methods, however, can be done even *before* model training which may not be holistically predictive of features that trained machine learning models would eventually find useful (Kumar and Chong, 2018). In addition, these methods break down as they only measure the linear relationship of linguistic variables in contrast to a possibility of a non-linear relationship of features in the dataset.

In order to have a clear understanding of how a

linguistic predictor affects model inference in readability assessment, the learned model weights of the said machine learning model should be extracted and analyzed *after* training. In this way, one can survey and rank the features used by a model and that has contributed substantially towards obtaining the final output. Thus, we lay down our contributions for this study as follows:

1. Feature selection through global interpretation of the state-of-the-art trained model in readability assessment for Filipino;

2. Analysis of the performance of readability assessment models trained using top features from global interpretation against top features from a correlation method; and,

3. Close-up analysis of top local features used by each grade level for readability analysis and their local weight boundaries through local surrogate interpretation.

2 Task Definition

We define the readability assessment task as a supervised learning problem. Given a document or specifically, a reading material d, a feature vector $x = [x_1, x_2 \ldots, x_n]$ is extracted and a model M is trained using said collection of linguistic predictors X along with the gold label Y or expert-identified readability level. From the trained model M, top n features $g = [g_1, g_2 \ldots, g_n]$ can be extracted through global interpretation and value boundaries (a, b) where $a, b \in \mathbb{R}$ of said features g_i can be identified through local surrogate interpretation.

3 Readability Assessment Models for Filipino

We survey different machine learning-based readability assessment models from previous works done for the Filipino language. We highlight the advantages and disadvantages of each model based on the data they used, scope of linguistic features, method of model training, and performance via select evaluation metrics.

Guevarra (2011). This is the first ever work that utilized the supervised machine learning methodology for readability assessment in Filipino. This work used logistic regression as the primary algorithm for model development and used 7 linguistic properties spanning frequency of words, unique words, sentences, average syllable count, word occurrence, log of word frequency, and percentage of top words. For the data, 140 children's books from Adarna House[1], the largest children's book publisher in the Philippines, were used. A relatively small error-rate was obtained using this approach during model evaluation with a score of 0.83 from the original grade level.

Bosque et al. (2011). This work pioneered the use of simple term frequencies and term-frequency inverse document frequency (TF-IDF), an NLP-based feature which represents each instance of the corpus as a vector of counts where dimensions are the words in the vocabulary. The data used are 105 essays and story books also from Adarna House and from the University of the Philippines Integrated School (UPIS). In this work, the TF-IDF representations are decomposed to reduce the sparsity of inputs. Readability assessment was viewed as a clustering task using Latent Semantic Indexing and Content Indexing as primary algorithms of choice for model development. The highest performance obtained is an average overall accuracy of 80.75% using Content Indexing with Term Frequency. One weakness that we highlight from this approach is that there are no linguistically-motivated features used such as variables leveraging on the semantics and syntactic structure of sentences.

Imperial et al. (2019). Improving the work of Bosque et al. (2011), this work explores wider NLP-based features including word n-grams and character-level n-grams. Similar dataset from Adarna House was used composed of 258 picture books converted to texts. For model development, this study explored on more complex ensemble architecture by combining K-Nearest Neighbors, Random Forest, and Multinomial Naive Bayes through a voting mechanism. Results showed a high performance of 82.2% using soft voting scheme. One weakness of the study is the

[1] https://adarna.com.ph/

use of heavily imbalanced dataset during training.

Imperial and Ong (2020a). This work pioneered the inclusion of true linguistically-motivated features as prescribed by experts (Macahilig, 2015) such as lexical features (adapted from Imperial and Ong (2020b)) capturing semantics and language model features capturing structure of sentences for a total of 25 predictors overall. For the data, leveled reading materials such as story books and activity books in the first three grade levels of basic education in the Philippines were used. For model development, the study made use of Logistic Regression and Support Vector Machines. Results showed that the inclusion of both lexical and language model features significantly increased the performance of readability assessment models by as high as 25%-32% in accuracy (from 33% to 72.0% for the top model).

Imperial and Ong (2021). Building from Imperial and Ong (2020a), this work explored even deeper linguistic features in Filipino spanning morphological features based on verb inflection and orthographical features based on syllable pattern with a total of 54 predictors overall. For the data, 265 reading materials from Adarna House (more than the study of Guevarra (2011)) and DepEd Commons[2], an online and open-source repository of children's resources in Filipino spanning first three grade levels, were used. For model training, the study explored Logistic Regression, Support Vector Machines, and Random Forest. From the 54 linguistic features used, the optimal subset of predictors was the combination of traditional and syllable pattern features obtaining a 66.1% accuracy using Random Forest. Although relatively lower than in the other previous works, this study empirically showed the learning process of the models as more features are used.

From the notable works listed in line with the development of readability assessment models for Filipino, we specifically chose to interpret the models trained by Imperial and Ong (2021). We emphasize that this work covers the most number of linguistic features used for any study and is the current state-of-the-art in the context of the Filipino language. Further details of replication and model interpretation are described in the succeeding sections.

4 Replication Setup

To interpret relations of model feature to its corresponding output, a trained model is needed first. We obtained such resource directly from the work of Imperial and Ong (2021).

4.1 Corpus

The study sourced children's books and reading materials from also Adarna House and DepEd Commons. DepEd Commons[3] is an online platform launched by the Department of Education (DepEd) of the Philippines. A total of 174 children's fictional books and 91 reading passages were used from both sources respectively. Both datasets have the same granularity of reading levels which are levels 1, 2, and 3 or L1, L2, and L3.

4.2 Linguistic Features

The study of Imperial and Ong (2021) extracted 54 different linguistic predictors spanning surface-based, lexical, language model, syllable structure, and morphological features—the most extensive study on the Filipino language to date. These features were also highlighted by past works that should be considered for the readability assessment task (Macahilig, 2015; Gutierrez, 2015). We briefly describe the components of each feature set:

Traditional or Surface-Based Features (TRAD). Word, sentence, phrase counts. Average word length, sentence length, and syllable count per word.

Lexical Features (LEX). Type-token ratio (TTR) and its variations. Noun and verb token ratio. Lexical density. Foreign and compound word density.

Language Model Features (LM). Perplexity scores of documents using language models trained from L1, L2, and L3 books in variations of unigram, bigram, and trigram splits.

[2]https://commons.deped.gov.ph/

[3]https://commons.deped.gov.ph/

Syllable Pattern Features (SYLL). Weighted frequencies of words of a document from the prescribed syllable patterns (*v, cv, vc, vcc, cvc, ccvc, ccvcc, ccvccc*) of the Philippine orthography.

Morphological Features (MORPH). Weighted frequencies of classified words based on morphology, specifically on verb inflection such as focus, aspect, and mood.

4.3 Model Performance

We trained the models using the same machine learning algorithms from the study of Imperial and Ong (2021) which were Logistic Regression, Support Vector Machines, and Random Forest. Likewise, we also obtained and set hyperparameter values as used by the study to obtain near similar replication of results. Table 1 shows the best scoring models using optimal subsets of features. For referencing purposes, the top feature sets for Logistic Regression are TRAD + LEX + SYLL + LM, TRAD + LM for Support Vector Machines, and TRAD + SYLL for Random Forest. These optimized models are highlighted in green in the table.

Model	Acc	Prec	Rec	F1
LogReg + All	0.542	0.530	0.542	0.532
LogReg + Best	**0.576**	**0.544**	**0.576**	**0.561**
SVM + All	0.492	0.481	0.492	0.485
SVM + Best	**0.525**	**0.524**	**0.525**	**0.524**
RF + All	0.627	0.623	0.627	0.624
RF + Best	**0.661**	**0.651**	**0.661**	**0.640**

Table 1: Comparison of top-ranked Spearman correlated features against top-ranked features via global interpretation on a Random Forest model.

4.4 Top Correlated Features

In addition to model training, we performed dependence analysis using Spearman correlation to identify if there are predictors from the linguistic features that correlate significantly with the readability levels. Table 2 shows the top 10 Spearman correlated features. From the table, majority of the feature sets are from TRAD and SYLL with a few from LEX and LM. The Type Token Ratio (TTR) emerged as the top negatively correlated feature relative to the task; however, a -0.3379 correlation value does not convey strong relationship by standard.

Feature Set	Predictor	Spearman's ρ
LEX	TTR	-0.3379
TRAD	polysyll words	0.3338
	average sentence len	0.3297
	word count	0.2745
LEX	BiLogTTR	-0.2723
SYLL	ccvc density	0.2620
	cvc density	0.2300
LM	L1 trigram	0.2247
SYLL	cvcc density	0.2075
	v density	0.1960

Table 2: Top 10 ranked Spearman's ρ scores for each linguistic feature.

5 Global Model Interpretation

The first part of model interpretation process is the global model interpretability of the readability assessment models trained using Logistic Regression, Support Vector Machines, and Random Forest. Global interpretation is the process of understanding the entire model as a blackbox by looking at the model's **learned global weights** based on features or possible predictors, in this case, the linguistically-motivated features for readability. Using the ELI5[4] package, the interactions of the best combination of the linguistic predictors producing the highest performance from the metrics were obtained for each model. We describe its implications and our findings in this section.

5.1 Logistic Regression Results

From the learned weights of the model in Figure 1 using the optimal feature subset of the linguistic predictors (TRAD + LEX + SYLL + LM), cross-referencing all features for all grade levels identified that CVC density score and L3 bigram feature were present across all levels. Thus, these two features can be used for readability assessment for Grades 1, 2 and 3. Meanwhile, the top unique features for each grade level are: noun-token ratio and L3 unigram for Grade 1; verb-token ratio, type-token ratio, compound word denstiy, and L3 trigram for Grade 2; and L1 bigram and root type-token ratio for Grade 3.

[4]https://eli5.readthedocs.io/en/latest/overview.html

y=0 top features		y=1 top features		y=2 top features	
Weight?	Feature	Weight?	Feature	Weight?	Feature
+1.083	L3_bigram	+0.806	L1_trigram	+0.802	L1_bigram
+0.646	BiLogTTR	+0.664	Verb_Token_Ratio	+0.581	cvc_density
+0.568	ccvcc_density	+0.491	cvc_density	+0.503	average_syllable_count
... 7 more positive ...		+0.382	L3_bigram	+0.490	polysyll_count
... 7 more negative ...		+0.313	vc_density	+0.448	RTTR
-0.664	polysyll_count	*... 5 more positive ...*		*... 13 more positive ...*	
-0.666	Noun_Token_Ratio	*... 9 more negative ...*		*... 6 more negative ...*	
-0.676	<BIAS>	-0.216	TTR	-0.483	ccvcc_density
-0.761	L1_trigram	-0.345	Compound_Word	-0.622	vc_density
-0.905	L3_unigram	-0.427	average_syllable_count	-0.679	BiLogTTR
-1.202	cvc_density	-0.648	<BIAS>	-1.015	<BIAS>
-1.366	average_sentence_len	-0.787	L3_trigram	-1.314	L3_bigram

Figure 1: Top 10 highly-important predictors for the Logistic Regression model using all linguistics features from the optimal subset. The green gradient indicates positive relationship while red indicates the opposite.

Focusing on the top feature with the highest weights for each grade level, for Grade 1 ($y = 0$), the negative weighted average sentence length is the most useful feature for helping the Logistic Regression model classify books as Grade 1. This may mean that ***as the average length value of sentences decreases, the readability level of texts may increase, proving that readability might not ultimately depend on a direct relationship with sentence length***.

On the other hand, for Grade 2 ($y = 1$), the combination of trigrams from the external Grade 1 activity books proved to be effective in helping the model classify story books to its category. For Grade 3 ($y = 2$), the same conclusion can be drawn regarding the inverse relationship of bigrams from external L3 activity books.

5.2 Support Vector Machine Results

From the learned weights of the model in Figure 2 using the optimal feature subset of the linguistic predictors (TRAD + LM), cross-referencing all features for all grade levels identified that L1 trigram, L3 unigram, and L3 bigram features were present across all grade levels. Thus, these three features can be used for readability assessment for Grades 1, 2 and 3. While these omnipresent features belong to the LM feature set, the TRAD features used were relatively unique for each grade level.

For Grade 1 ($y = 0$), the most useful TRAD features are word count observing positive relationship with respect to the readability levels, and average syllable and sentence counts observing negative relationship with respect to the readability levels. For Grade 2 ($y = 1$), the most useful TRAD features are average word length and polysyllable count observing positive and negative relationships with the readability levels respectively. For Grade 3 ($y = 2$), the most useful TRAD features are average sentence length, average syllable count, sentence count, and total polysyllable words all of which observe positive relationships with the readability levels.

5.3 Random Forest Results

For the Random Forest model weights as described in Figure 3, the top five features all belong to the TRAD feature set followed by features from SYLL. The simple feature of raw word counts (WC) observing positive relationship with the readability levels emerged as the highest scoring feature. Thus, ***the higher the word count of a story book, the higher probability of the story book belonging to a higher readability level***.

Testing for significant difference[5] in the word count of the data from Grade 1 against Grades 2 and 3 obtained p values of 0.004 and 0.014, respectively. Thus, the word count of story books from the Grade 1 dataset is *significantly* different compared to Grades 2 and 3, proving the initial hypothesis.

Interestingly, all of the top TRAD features used

[5]Two sample t test was conducted over $\alpha = 0.05$.

y=0 top features		y=1 top features		y=2 top features	
Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature
+0.194	actor_focus_ratio	+0.185	Verb_Token_Ratio	+0.129	polysyll_count
+0.131	BiLogTTR	+0.114	vc_density	+0.090	Lexical_Density
+0.108	L3_bigram	+0.111	cvc_density	+0.080	average_sentence_len
... 14 more positive ...		+0.082	L1_unigram	*... 22 more positive ...*	
... 27 more negative ...		*... 21 more positive ...*		*... 19 more negative ...*	
-0.105	L1_trigram	*... 22 more negative ...*		-0.085	L2_unigram
-0.106	polysyll_count	-0.074	phrase_count_per_sentence	-0.092	ccvcc_density
-0.113	word_count	-0.077	TTR	-0.108	actor_focus_ratio
-0.149	Verb_Token_Ratio	-0.104	polysyll_count	-0.115	TTR
-0.166	cvc_density	-0.111	Compound_Word	-0.128	BiLogTTR
-0.176	<BIAS>	-0.139	actor_focus_ratio	-0.146	L3_bigram
-0.218	average_sentence_len	-0.215	<BIAS>	-0.184	<BIAS>

Figure 2: Top 10 highly-important predictors for the Support Vector Machine model using all linguistics features from the optimal subset. The green gradient indicates positive relationship while red indicates the opposite.

Weight	Feature
0.1022 ± 0.1213	word_count
0.0963 ± 0.1057	average_sentence_len
0.0900 ± 0.1164	polysyll_count
0.0702 ± 0.0964	sentence_count
0.0658 ± 0.0900	phrase_count_per_sentence
0.0602 ± 0.0866	cvc_density
0.0567 ± 0.0814	vc_density
0.0526 ± 0.0751	ccvc_density
0.0518 ± 0.0814	average_word_len
0.0492 ± 0.0730	ccvcc_density
... 7 more ...	

Figure 3: Top 10 highly-important predictors for the Random Forest model using all linguistics features from the optimal subset. All features observe positive relationship with the readability levels.

by the best Random Forest model are present in traditional, formula-based readability indices. For example, in the formula developed by Villamin (1979), average sentence length was included; in Guevarra (2011), word count, sentence count, and polysyllable words were included; and in Macahilig (2015), word count was included. One important inference from this observation is that ***the use of TRAD features is still practical for readability assessment of Filipino texts, specifically for children's story books***. However, other deeper linguistic feature sets, such as language model features, can be integrated to these TRAD predictors to produce models capable of achieving better performance for the task.

5.4 Global Features vs. Correlated Features

Table 3 compares the performance of running Random Forest models using the top 10 Spearman correlated features from the study of Imperial and Ong (2021) as reported in Table 2 against using the top 10 features reported in Figure 3 using global interpretation. From the results, using top weighted features through global interpretation of model obtained much higher performance in contrast to using Spearman features. To be more specific, the increase in scores are 5% for accuracy, 6.4% for precision, 7.3% for recall, and 13.7% for F1.

It is also worth mentioning that there are coinciding linguistic features present in both the top Spearman and global interpretation lists such as word count, polysyllable count, and average sentence length. Combining the 15 unique features from both lists and retraining the Random Forest model produces an even higher performance score with 69.8% in accuracy, ≈19% increase from using global features. With this result, we infer that ***using top features obtained via global interpretation combined with top correlated Spearman features achieves a much higher performance than using correlated features alone*** as done in previous works.

6 Local Surrogate Interpretation

The second part of the model interpretation process is the local surrogate interpretability. While global interpretation takes a look at the overall model's

Model	Acc	Prec	Rec	F1
RF + Corr	0.458	0.464	0.435	0.432
RF + Global	0.508	0.528	0.508	0.512
RF + Combined	**0.698**	**0.682**	**0.624**	**0.649**

Table 3: Comparison of top-ranked Spearman correlated features against top-ranked features via global interpretation on a Random Forest model.

learned weights, local surrogate interpretation of models focuses on the local analysis for a particular data point (Yang et al., 2018). We select the best-performing model for this experiment which is the Random Forest model using TRAD + SYLL features as described in Table 1. Specifically, the local surrogate interpretation technique used in this experiment is the extraction of **learned local weight boundaries** used by random tree estimators for correctly classifying different instances from the data of three grade levels. This process was done using the LIME package[6] developed in Ribeiro et al (2016).

As seen in Figure 4, **the average sentence length is a consistent predictor across all readability levels** used by the best Random Forest model as it is present within the top 10 features. In the case of the test instance for Grade 1, the average sentence length boundary value is $<= -0.43$. Thus, increasing this value would mean that the model may count it towards the other grade levels. For the test instance of Grade 2, the boundary value for average sentence length is $-0.49 < X <= -0.01$, observing a larger range than in the test instance of Grade 1. Thus, if the feature weight of average sentence length falls within this boundary, the model will classify the instance as Grade 2. Similarly, the boundary value of average sentence length for Grade 3 is $-0.04 < X <= 0.55$ wherein if the average weight of a test instance falls within this range, the model will likely label it in favor of Grade 3.

To emphasize, the *connecting* local weight boundaries of linguistic predictors can be traced from Grade 1 to Grade 3 by referencing the values shown in Figure 4. For example, the boundary value of polysyllable count (frequency of words with more than five syllables) for Grade 1 is $<= -0.33$, for Grade 2 this becomes $-0.33 < X <= 0.00$, and for Grade 3 this becomes > 0.50 indicating that ***a reading material with a high count of polysyllable words might signal the model to attribute it to a higher readability level***.

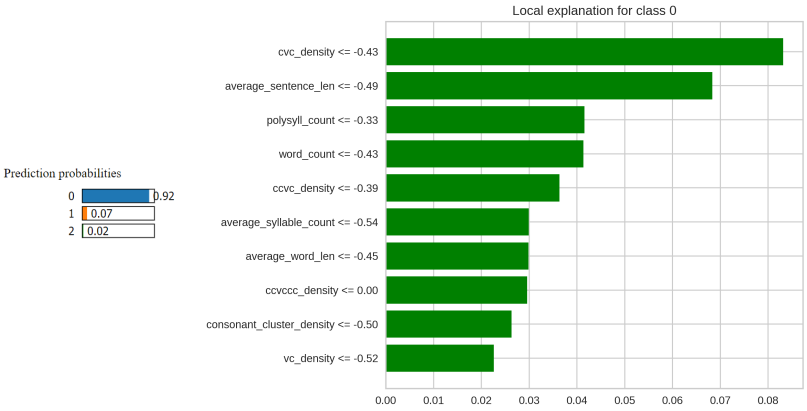
It is important to note that these learned boundaries from the best-performing models are derived only from random instances of the test set. This experiment was performed to determine the conditions that were being used by the learned model to help it discriminate between readability levels. It should also be emphasized that the model will not instantly classify an instance if only one feature or predictor falls within or satisfies the boundary values. Thus, ***majority of the weights of the top linguistic predictors of a test instance should satisfy the accumulated boundary conditions for it to be classified to a certain category***.

7 Conclusion

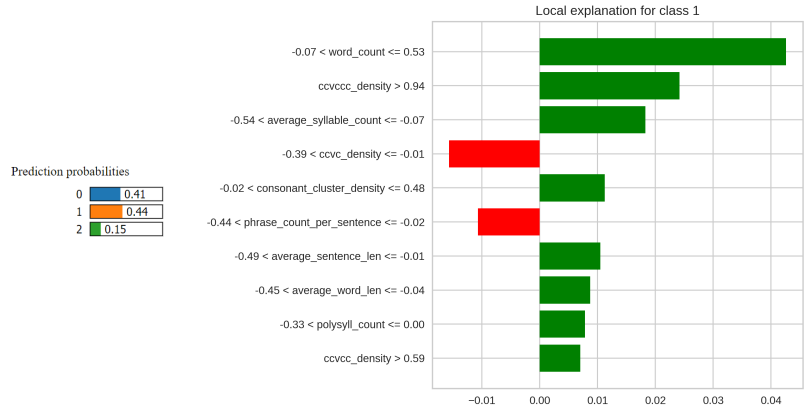
Readability assessment is the process of gauging a certain piece of text or reading material in terms of reading difficulty. In this study, we addressed one of the major challenges in readability analysis by exploring feature interpretation using global and local surrogate methods in the context of the Filipino language. Results showed that using global interpretation as a feature selection technique to extract top contributory linguistic features obtained higher performance across all metrics than using the standard correlation method. In addition, combining these two processes would further increase the performance of the model in readability assessment tasks. We also empirically observed the local boundaries of the top-performing model for each linguistic feature. From these boundaries, we learned how perturbing values can directly affect the final readability classification of a certain book or reading material. Future directions of this study include interpretation of complex neural models for readability assessment as there is an increase of patronization in this direction from the research community. Likewise, stronger causality measures such as Bayesian networks can be explored to ground hypotheses on cause and effects of perturbation of local weight boundaries of features.

Acknowledgment The authors would like to thank the anonymous reviewers for their valuable

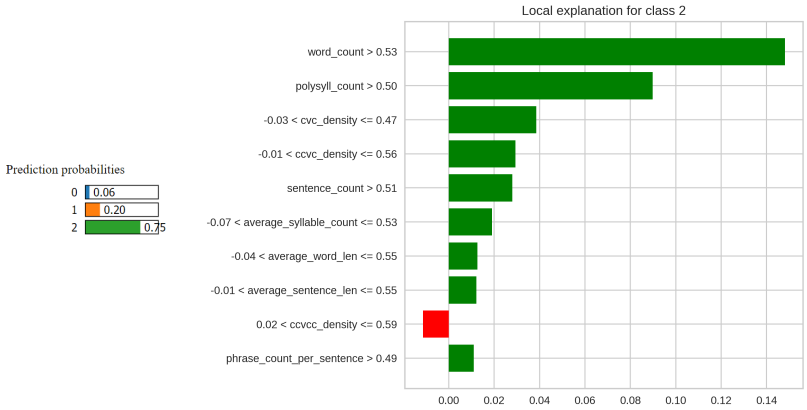
[6] https://github.com/marcotcr/lime



(a) Grade 1



(b) Grade 2



(c) Grade 3

Figure 4: Local explanations of random test instances for each grade level using the best-performing Random Forest model. Green bars indicates positive relationship while red indicates the opposite.

feedback and to Dr. Ani Almario of Adarna House for allowing us to use their children's book dataset for this study. This work is also supported by the DOST National Research Council of the Philippines (NRCP).

References

[Chatzipanagiotidis et al.2021] Savvas Chatzipanagiotidis, Maria Giagkou, and Detmar Meurers. 2021. Broad linguistic complexity analysis for Greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58, Online, April. Association for Computational Linguistics.

[Dale and Chall1948] Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

[Deutsch et al.2020] Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

[Guevarra2011] Rowena C. Guevarra. 2011. Development of a Filipino text readability index.

[Gutierrez2015] Merry Ruth M Gutierrez. 2015. The suitability of the fry and smog readability formulae in determining the readability of filipino texts. *The Normal Lights*, 8(1).

[Hancke et al.2012] Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

[Imperial and Ong2020a] Joseph Marvin Imperial and Ethel Ong. 2020a. Exploring hybrid linguistic feature sets to measure filipino text readability. In *2020 International Conference on Asian Language Processing (IALP)*, pages 175–180. IEEE.

[Imperial and Ong2020b] Joseph Marvin R Imperial and Ethel C Ong. 2020b. Application of lexical features towards improvement of filipino readability identification of children's literature.

[Imperial and Ong2021] Joseph Marvin Imperial and Ethel Ong. 2021. Diverse linguistic features for assessing reading difficulty of educational filipino texts. *arXiv preprint arXiv:2108.00241*.

[Imperial et al.2019] Joseph Marvin Imperial, Rachel Edita Roxas, Erica Mae Campos, Jemelee Oandasan, Reyniel Caraballo, Ferry Winsley Sabdani, and Ani Rosa Almaroi. 2019. Developing a machine learning-based grade level classifier for filipino children's literature. In *2019 International Conference on Asian Language Processing (IALP)*, pages 413–418. IEEE.

[Kincaid et al.1975] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

[Kumar and Chong2018] Sunil Kumar and Ilyoung Chong. 2018. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *International journal of environmental research and public health*, 15(12):2907.

[Macahilig2015] Heidi B. Macahilig. 2015. A content-based readability formula for Filipino texts. *The Normal Lights*, 8(1).

[Razon et al.2011] Abigail R Razon, Rommel R Ledesma, Michelle Louise S Bosque, Hazel M Loberas, Ani Rosa S Almario, Rosalie C Faune, Rowena Cristina L Guevara, and Prospero C Naval Jr. 2011. Readability analysis of grade school reading books using concept indexing with k-means clusteringm. *International Symposium on Multimedia and Communication Technology*.

[Reynolds2016] Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA, June. Association for Computational Linguistics.

[Ribeiro et al.2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

[Vajjala and Meurers2012] Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

[Vajjala2021] Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *CoRR*, abs/2105.00973.

[Villamin and de Guzman1979] Aracelli M. Villamin and E.S. de Guzman. 1979. Pilipino readability formula: The derivation of a readability formula and a Pilipino word list.

[Weiß and Meurers2018] Zarah Weiß and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

[Xia et al.2016] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA, June. Association for Computational Linguistics.

[Yang et al.2018] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. 2018. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570. IEEE.