A study of Vietnamese readability assessing through semantic and statistical features

Hung Tuan Le[1,2], Long Truong To[1,2], Manh Trong Nguyen[1,2], Trong-Hop Do[1,2*]

[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): hopdt@uit.edu.vn;
Contributing authors: 21520250@gm.uit.edu.vn; 21521101@gm.uit.edu.vn; 21520343@gm.uit.edu.vn;

Abstract

The difficulty of a text is a problem that involves assessing language factors that impact a person's reading comprehension ability. However, current research in Vietnamese has only focused on statistical features. This paper introduces a new approach that integrates statistical and semantic methodologies to evaluate text difficulty. Our research utilizes three distinct datasets: the Vietnamese Text Readability Dataset (ViRead), OneStopEnglish, and RACE, translated into Vietnamese. Advanced semantic analysis methods are employed for the semantic aspect using state-of-the-art language models such as PhoBERT, ViDeBERTa, and ViBERT. In contrast, statistical methods are incorporated to extract syntactic and lexical features. We conduct experiments using various machine learning models, including Support Vector Machine (SVM), Random Forest, and Extra Trees and evaluate their performance using accuracy and F1 score metrics. Our results indicate that the joint approach of combining semantic and statistical features significantly enhances the accuracy of readability classification compared to using each method in isolation. The research emphasizes the importance of considering both linguistic and semantic aspects for a more accurate assessment of text complexity in Vietnamese. This contribution to the field provides insights into the adaptability of advanced language models in the context of Vietnamese text readability. It lays the groundwork for future research in this area.

Keywords: Vietnamese Text Readability, Semantic Analysis, Statistical Approach, Language Models, Machine Learning, Text Complexity Assessment.

1 Introduction

Exchanging information and knowledge through text has led to the emergence of assessing text difficulty. There can be multiple ways to describe and convey content when dealing with the same issue. Among them, complex texts pose challenges for readers, as reflected in reading speed, comprehension, and the ability to connect information within the text. In recent years, text difficulty has been evaluated through features present, such as syntactic complexity, complexity in logical relationships and inferences of information in the text, and the sequential expression of data over time or context.

The methods for assessing text difficulty can be classified into two main approaches: statistical and machine learning or deep learning. In the statistical approach, text difficulty is evaluated through the synthesis of easy-to-compute features in the text, such as the length of the text, the average number of words and sentences in the text, etc. Kincaid et al., Flesch, where these features are extracted and evaluated through correlation analysis with the difficulty of a set of texts. The machine or deep learning approach involves using neural models to represent the semantics present in the text, allowing for the assessment of text difficulty [3–6].

Studies addressing the problem by applying modern neural models such as BERT and its variants combined with features extracted through traditional statistical methods have achieved promising results on English datasets such as WeeBit [7], OneStopEnglish [8], and Cambridge [9]. In Vietnam, pioneering research in solving the problem, such as that of [10, 11], and more recently [12], has applied PhoBERT - a pre-trained language model [13] designed explicitly for Vietnamese - to address the problem. However, these studies assess text difficulty through individual sentences, overlooking features between sentences, such as discourse relations, and entity cohesion.

Identifying the shortcomings in studies related to both semantic and statistical features in Vietnamese readability research, in today's paper, we scrutinize the impacts of statistical and semantic features, as well as the correlation between these two types of features on the difficulty of Vietnamese texts across three primary datasets: Vietnamese Readability dataset [14], RACE [15], and OneStopEnglish [8]. Our methods range from traditional machine learning models such as SVM, Random Forest, and Extra Tree to state-of-the-art pre-trained language models in various semantic tasks, such as PhoBERT [13], ViDeBERTa [16], and ViBERT [17]. The results show that statistical and semantic features improve model performance, although not yet surpassing statistical features alone. However, they demonstrate potential for development on larger datasets.

Furthermore, we conduct an in-depth analysis of specific groups of statistical features concerning text difficulty by individually examining each feature group across multiple models. The results show that features such as 'Number of words' or 'Average word length in characters' have the most significant impact on the models when combined with semantic features from deep learning models.

2 Related Work

In this section, we will survey existing research on the readability problem globally (see section 2.1) and current research outcomes in Vietnamese (see section 2.2).

2.1 Textual Readability

Research on this problem worldwide has gained attention from the research community in the field of natural language processing, especially the early studies in English such as [2], approaching the problem statistically by examining the readability of text through the number of syllables/words. Later, in 1975, the readability index by Kincaid et al. was published based on the features of [2]. In Chall and Dale, the readability of the text was assessed based on the semantic difficulty of words in the text by examining the frequency of word occurrences with a word list of 3000 words. In the following years, these features became standards for evaluation [19, 20], along with syntactic features such as the height of the parse tree [18]. However, the statistical approach still cannot extract more deep linguistic features that directly affect the readability of the text, such as discourse relations, cohesion, or rhetorical organization [21].

With the development of language models and the increase in the amount of training data, a new approach to the Readability problem has emerged by leveraging the language representation capabilities of these models to extract more deep linguistic features and the classification ability of probabilistic models or deep learning models. Early studies include those by Si and Callan and Collins-Thompson and Callan who applied unigram language models and classification through naive Bayes. In the following years, the probabilistic model approach gained attention and achieved good results [4, 5, 23, 24]. After the emergence of deep learning models, especially the development of pre-trained language models with transformer architecture, which achieved state-of-the-art results on other semantic tasks, performance on the readability problem improved by feature extraction through models [25–27] and also by combining with other externally collected statistical features [3, 28, 29].

In addition to English, research on other languages has also been developed based on the existing research foundation in English, such as for French [30], Italian [31], German [32], Swedish [33, 34], Bangla [35], and Greek [36].

2.2 Vietnamese Readability

Research on the readability problem still needs to be improved due to the lack of quality datasets, as seen in studies from [10, 37] to [11, 14, 38], where the dataset sizes are petite (less than 2000 samples). Additionally, the primary approach to the problem involves extracting features through statistical analysis, such as the number of syllables/words, height/width of parse tree, and number of clauses [39]. Recently, Doan et al. approached the problem from a different perspective by extracting features through PhoBERT [13]; however, the research is currently unavailable to the community.

3 Vietnamese Readability

In this section, we describe the experimental process in the paper, including the datasets (see section 3.1) and the methods we experimented with (see section 3.2).

3.1 Datasets

We use a total of 3 datasets described in Table 1, namely OneStopEnglish [8], RACE [15], and the Vietnamese Text Readability Dataset [14].

The Vietnamese Text Readability Dataset (ViRead) [14] is constructed from Vietnamese college-level textbooks, stories, and literature websites. After extracting text from these sources through OCR, sentences are labeled by a team of twenty teachers who teach Vietnamese literature at middle and high schools and colleges. The labels have four categories: Very Easy, Easy, Medium, and Difficult.

Due to the lack of large-scale and high-quality datasets in Vietnamese for the readability problem, we also use two English datasets: OneStopEnglish [8] and RACE [15]. The OneStopEnglish dataset is extracted from onestopenglish[1], an English language learning resources website run by MacMillan Education. The content is rewritten in three versions from The Guardian newspaper, with three versions labeled as advanced (Adv), intermediate (Int), and elementary (Ele). The RACE dataset is a large-scale Reading Comprehension benchmark extracted from English exams for Chinese students from middle school and high school, including 28,000 passages. In the readability task, RACE is divided into junior and senior.

We translated the two English datasets, OneStopEnglish and RACE, into Vietnamese using Google Translate[2]. Subsequently, we split these datasets into smaller components for the experimentation process. Due to the small size of the OneStopEnglish and ViRead datasets (under 2000 samples), we divided them into two sets: a training set (train) and a test set (test). The size statistics for each dataset are provided in Table 1.

Datasets	Domain	Language	Number of sample	Number of class	Training	Test
ViREAD	Literature	Vietnamese	1825	4	1460	365
Race	Education	English	27933	2	22346	5587
OneStopEnglish	Educaion	English	567	3	453	114

Table 1: Datasets statistics

3.2 Empirical Method

In this section, we proceed to design the implementation process along two main approaches: the statistical approach (see section 3.2.1) and the semantic approach (see section 3.2.2). The statistical approach involves using statistical methods to extract features from the dataset. In contrast, the semantic approach utilizes machine learning models, from simple to deep learning models, to extract semantic features from the dataset. Additionally, we conduct experiments combining features from statistical and semantic approaches to understand the correlation between them and the results (see section 3.2.3).

3.2.1 Statistical approach

Luong et al. conducted experiments on the influence of features on the readability of texts extracted in a statistical approach on the Vietnamese readability dataset [14]. The features considered include part-of-speech features (ratio of POS tag words, common nouns/distinct words), syntax-level features (average tree depths (parse tree)), and Vietnamese-specific features (ratio of borrowed words, ratio of Sino-Vietnamese words). We utilized features with high correlation with the difficulty of the text, and these features are listed in table 2.

[1] https://onestopenglish.com/
[2] https://translate.google.com/

In addition, we incorporated two new features, such as word cohesion represented through dependency trees, to understand how relationships between word cohesion in a sentence influence the difficulty of the text (see table 2). To extract these two features, we employed VnCoreNLP [40] for sentence segmentation and subsequent dependency representation. The statistical features will be classified by three machine learning models: Support Vector Machine (SVM), Random Forest, and Extra Trees.

Category	Feature
Raw Feature	Number of words
	Average word length in character
	Ratio of long sentence (in syllable)
POS Feature	Distinct common nouns/distinct words
	Distinct parallel conjunctions/distinct words
	Ratio of single POS tag words
	Adverbs/sentences
Syntas-Level Feature	Average no. distinct conjunction word
	Average no. conjunction word
Vietnamese-Specific Feature	Ratio of borrowed words
	Ratio of distinct borrowed words
	Ratio of distinct Sino-Vietnamese words
Word Cohension	Depth of Dependency Tree
	Average overlapping between multiple sentences in paragraph

Table 2: Linguistic features

The statistical features on the three datasets ViRead, OneStopEnglish, and RACE are summarized in Table 3. As observed, in translated datasets such as OneStopEnglish and RACE, some standard features in the text remain similar, such as 'Average word length in character' and 'Distinct parallel conjunctions/distinct words.' For 'Ratio of long sentence' features, we choose sentences with more than 20 syllables according to a research of American Press Institute. However, specific features for Vietnamese, such as the 'Ratio of borrowed words' or 'Ratio of distinct Sino-Vietnamese words,' differ. This discrepancy is due to the translation and certain unique features in Vietnamese texts. These differences' impact directly influences the models' results, as discussed in Section 4.

3.2.2 Semantic approach

In this section, we employ advanced semantic analysis methods for classifying the difficulty level of Vietnamese texts. Our semantic approach primarily utilizes three state-of-the-art language models: PhoBERT [13], ViDeBERTa [16], and ViBERT [17]. These models are instrumental in extracting deep semantic features from the Vietnamese texts, which are crucial for our classification task.

PhoBERT [13] emerges as a paragon, trained extensively on a corpus comprising 20GB of Vietnamese Wikipedia and news text. It boasts 135 million parameters in its base iteration and an augmented 370 million parameters for the large variant. In its most recent iteration,

Feature	ViRead	OneStopEnglish	RACE
Number of words	40 - 23104	263 - 1417	2 - 1271
Average word length in character	2.4973 - 3.4071	2.9754 - 3.501792	2.287 - 5.483
Ratio of long sentence (in syllable)	0 - 1	0.2714 - 1	0 - 1
Distinct common nouns/distinct words	0.0312 - 0.44	0.1194 - 0.2612	0 - 0.5
Distinct parallel conjunctions/distinct words	0- 0.1129	0.0052 - 0.0284	0 - 0.1739
Ratio of single POS tag words	0.7977 - 1	0.8815 - 0.9627	0.8421 - 1
Adverbs/sentences	1 - 82	7 - 34	0 - 39
Average no. distinct conjunction word	0 -36	3 -18	0 -18
Average no. conjunction word	0 -1670	11 - 77	0 - 79
Ratio of borrowed words	0 - 0.0128	0 - 0.0279	0 - 0.0058
Ratio of distinct borrowed words	0 - 0.0085	0 - 0.0085	0 - 0.044
Ratio of distinct Sino-Vietnamese words	0.0317 - 0.4179	0.0022 - 0.0149	0 - 0.396
Depth of Dependency Tree	1.5 - 30.3333	6.8966 - 21.1053	1 - 132
Average overlapping between multiple setence in paragraph	0.2539 - 143.2710	1.6590 - 10.5664	0 - 11.157

Table 3: The min-max extraction result of statistical features in ViRead, OneStopEnglish and RACE

PhoBERT$_{base} - V2$, the model has been refined on a formidable 120GB of Vietnamese text derived from the OSCAR-2301 dataset[3].

ViDeBERTa [16] is a model with the architecture of DeBERTa [41] and has been trained on CC100[4] corpus, including 138GB uncompressed texts. ViDeBERTa outperforms PhoBERT on tasks such as named entity recognition (NER) and part-of-speech (POS). However, the current version of ViDeBERTa with the DeBERTa-V3 architecture has not been released; instead, the version with the DeBERTa$_{Base}$-V2 architecture is available [5]. ViBERT [17] has been trained on approximately 10GB of texts collected from online newspapers in Vietnamese, enabling the model to represent the semantics of words more effectively.

The extracted features from the pre-trained language models will be classified using machine learning models such as Support Vector Machine (SVM), Random Forest, and Extra Trees, as well as deep learning models such as Multi-Layer Perceptron (MLP).

3.2.3 Joint approach

We investigate the synergy between statistical and semantic approaches by conducting experiments combining features from both methods. The experiment aims to comprehend the complementary nature of these approaches and how they can be seamlessly integrated to enhance the accuracy of difficulty classification. Features extracted through the methods in section 3.2.1 and section 3.2.2 will be concatenated and fed into classification models, including SVM, random forest, and extra tree.

3.2.4 Evaluation Metric

To assess the performance of the models in our experiments, we employ accuracy and F_1 score (macro average) as the two main evaluation metrics, where the F_1 score is described below:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

[3]https://huggingface.co/datasets/oscar-corpus/OSCAR-2301
[4]https://huggingface.co/datasets/cc100
[5]https://huggingface.co/Fsoft-AIC/videberta-base

4 Experiment Result

4.1 Statistical Result

Dataset	Model	Result	
		F1	**Acc**
ViRead	SVM	88.48	92.05
	Random Forest	**92.59**	**95.34**
	Extra Tree	91.34	94.52
OneStopEnglish	SVM	72.85	72.81
	Random Forest	74.97	74.56
	Extra Tree	**75.77**	**75.44**
RACE	SVM	71.27	76.67
	Random Forest	72.77	77.07
	Extra Tree	**72.84**	**77.07**

Table 4: Statistical approach performance on machine learning model

From the results described in Table 4, we observe that the Extra Tree model performs well on both OneStopEnglish and RACE datasets. Specifically, on the OneStopEnglish dataset, Extra Tree outperforms the other two models, SVM and Random Forest, being 0.8% higher than the second-best model (Random Forest) in terms of F_1-score and 2.92% higher than SVM in F_1-score. In the RACE dataset, Extra Tree remains the top-performing model. However, the performance gap between Extra Tree and the other two models is negligible, with a 0.07% difference with Random Forest and a 1.57% difference with SVM in terms of F_1-score. This discrepancy in performance between Extra Tree and the other models on OneStopEnglish and RACE datasets may be attributed to the substantial difference in dataset sizes, with OneStopEnglish having only 567 samples. In comparison, RACE has a larger dataset of 27,933 samples.

In contrast to the cases in the RACE and OneStopEnglish datasets, on the ViRead dataset, Random Forest is the top-performing model with an F_1-score of 92.58%, followed by Extra Tree with 91.34%, and SVM with 88.48%. The superior performance of the ViRead dataset can be attributed to the fact that RACE and OneStopEnglish datasets are translations from English to Vietnamese, resulting in fewer distinctive features unique to Vietnamese compared to ViRead, a dataset extracted from Vietnamese-language textbooks.

4.2 Semantic Result

The experimental results using the language representation capabilities of pre-trained language models are summarized in Table 5. The statistical results demonstrate that PhoBERT's semantic representation outperforms ViDeBERTa and ViBERT on the OneStopEnglish and RACE datasets, achieving a 63.66% F_1 score on the OneStopEnglish dataset and a 74.5% F_1 score on the RACE dataset when using MLP for classification. However, on the OneStopEnglish dataset, when employing other classification models such as Random Forest and Extra Tree, features

Semantic approach		Result							
		F1				Acc			
		MLP	**SVM**	**Random Forest**	**Extra Tree**	**MLP**	**SVM**	**Random Forest**	**Extra Tree**
ViRead	PhoBERT	72.45	64.43	79.17	77.4	80	80.55	83.56	84.66
	ViDeBERTa	44.45	14.84	76.34	**80.11**	59.73	42.19	81.92	**84.93**
	ViBERT	63.17	62.08	75.36	73.7	73.7	77.81	82.19	83.01
OneStopEnglish	PhoBERT	**63.66**	41	29.37	15.59	**64.91**	48.25	28.95	14.91
	ViDeBERTa	40.13	18.56	55.35	52.32	46.49	30.7	54.39	53.51
	ViBERT	41.45	31.02	32.78	19.66	42.98	37.72	33.33	20.18
Race	PhoBERT	**74.5**	72.96	71.82	70.67	**79.2**	77.89	76.64	76.52
	ViDeBERTa	60.16	56.69	66.22	64.9	70.93	70.28	72.1	72.12
	ViBERT	70.01	68.92	69.06	66.81	75.47	75.8	74.65	74.13

Table 5: Semantic approach using both pre-trained language models and machine learning model

extracted through PhoBERT yield lower results in both $F1_1$ score and accuracy compared to features extracted through ViDeBERTa. Nevertheless, when using SVM for classification, features extracted through PhoBERT outperform those extracted through ViDeBERTa. This discrepancy may be attributed to the small training dataset size in the OneStopEnglish dataset, leading to unusual model performance variations, unlike the RACE dataset where the performance of classification models using features extracted through PhoBERT consistently outperforms those using ViDeBERTa and ViBERT.

Similarly, the performance of classification models using features extracted through PhoBERT is generally higher than ViDeBERTa, except for one exceptional case when classifying with the Extra Tree model. In this case, the ViDeBERTa embeddings outperform PhoBERT embeddings by 2.71% in terms of F_1 score and 0.27% accuracy. This anomaly may be attributed to the small dataset size, leading to unclear and unstable differences between the two embedding methods.

Furthermore, significant variations in results are observed when comparing the performance of models determining difficulty through the semantic representation of pre-trained language models with conventional classification models using features derived from statistics. For instance, on the ViRead and OneStopEnglish datasets, the models with combined semantic and statistical features yield lower results than those employing only statistical features. This could be attributed to the limited size of the training data, causing a decrease in performance, contrary to the models trained on the RACE dataset. However, the RACE dataset needs more Vietnamese language features, resulting in only marginal performance improvement.

4.3 Joint Result

The experimental results of the classification models with the combination of features, including embeddings from pre-trained language models and statistical features, are summarized in Table 6. Overall, across the three datasets, the feature combination method significantly improves the performance of the models compared to using only features extracted by transformers (see section 4.2).

In the ViRead and OneStopEnglish datasets, the classification models' performance increases from 17.255% to over 37.01% in terms of F_1 score and from 11.3675% to 27.41% in terms of accuracy across the three different feature extraction methods. However, in the

Joint Approach		Result							
		F1				Acc			
		MLP	SVM	Random Forest	Extra Tree	MLP	SVM	Random Forest	Extra Tree
ViRead	PhoBERT	91.76	87.52	**92.17**	90.06	94.52	92.05	**94.52**	93.15
	ViDeBERTa	91.23	87.84	91.92	92.15	94.25	91.33	94.25	94.52
	ViBERT	86.2	86.37	90.82	89.35	91.51	90.11	93.7	92.33
OneStopEnglish	PhoBERT	67.96	72.66	56.26	45.2	69.3	73.68	56.14	45.61
	ViDeBERTa	67.29	**73.72**	64.91	64.51	70.18	**73.88**	64.35	64.91
	ViBERT	56.33	71.55	60.93	49.54	58.77	72.64	61.4	50
Race	PhoBERT	73.17	71.62	73.97	**77.09**	78.27	77.69	**78**	77.2
	ViDeBERTa	64.34	70.98	73.02	69.85	74.53	76.53	77.33	75.2
	ViBERT	71.27	71.19	72.46	71.07	77.6	76.67	76.67	76.43

Table 6: Joint approach result when combine both statistical and embedding features

RACE dataset, the performance improvement of the models is not substantial, only increasing by an average of 4% across all three embedding methods. Additionally, some cases show that the model's performance decreases when combining features, such as SVM and MLP, when extracted by PhoBERT. This is likely because the SVM and MLP models rely on certain Vietnamese-specific features that are less present in the RACE dataset than in the ViRead dataset.

Although the combined feature results are slightly lower than using only statistical features (see section 4.1)—lower by 0.42% in F_1 score and 0.82% in accuracy on the ViRead dataset, and 2.05% in F_1 score and 1.56% in accuracy on the OneStopEnglish dataset—the small size of these two datasets may contribute to this observation. If the dataset size is increased, as in the case of the RACE dataset, where combining features improves performance, then combining features is likely to lead to improvements in readability classification.

5 Experiment Analysis

Dataset	Model	Raw Feature		POS Feature		Syntax-Level Feature		Vietnamese-Specific Feature		Word Cohesion	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
ViRead	PhoBert + MLP	94.79	92.1	**95.07**	**92.83**	93.7	91.38	80	76.84	93.42	91
	PhoBERT + Random Forest	**93.7**	**90.7**	92.6	89.83	90.68	86.69	83.56	77	87.67	82.06
OneStopEnglish	PhoBert + MLP	56.14	46.06	56.14	55.03	44.74	36.8	57.02	54.76	**79.09**	**70.18**
	PhoBert + SVM	**72.81**	**72.78**	64.91	64.93	43.86	37.38	54.39	53.99	58.77	59.35
RACE	PhoBert + MLP	**78.75**	**75.35**	78.55	74.46	78.89	75.34	77.79	74.11	78.61	75.24
	PhoBERT + Extra Tree	**77.63**	**72.36**	76.78	71.01	76.96	71.21	76.56	70.63	76.7	70.86

Table 7: The effect of statistical features on the performance of the model when combining both Embedding and statistical features

We utilized the best-performing models on each dataset from Section 4.3 and further conducted individual experiments on each group of features, including statistical features and features obtained through pre-trained language models. The experimental results are summarized in Table 7.

Generally, the feature group that most influence the models when combining statistical and embedding features is the 'Raw Feature',' followed by 'POS Feature,' 'Word Cohesion', 'Syntax-Level Feature,' and finally the 'Vietnamese-Specific Feature'. The improvement in model performance when using the 'Raw Feature' group alone is understandable. This

Dataset	Model	Acc 25%	Acc 50%	Acc 75%
ViRead	PhoBERT + MLP	82.61	95.63	96.35
	Random Forest	98.91	99.45	98.18
	PhoBERT + Random Forest	92.39	97.81	97.45
OneStopEnglish	PhoBERT + MLP	37.93	54.39	65.88
	Extra Trees	86.21	75.44	80
	PhoBERT + SVM	86.21	68.42	57.65
RACE	PhoBERT + MLP	80.86	79.04	79.52
	Extra Trees	80.24	77.33	78.54
	PhoBERT + Extra Tree	78.8	77.65	77.77

Table 8: Accuracy of models according to data size

is because texts with many sentences and words per sentence encompass vast knowledge, directly influencing the text's difficulty by requiring readers to absorb a significant amount of information. Combining features from the 'Raw Feature' group with machine learning models significantly enhances the model's performance.

Apart from the 'Raw Feature' group, the 'POS Feature' and 'Word Cohesion' feature groups also affect the model's performance. In 'POS Feature,' if a text contains many polysemous words, the complexity of the text increases, requiring readers to understand the context of the sentence to truly comprehend the intended meaning of the ambiguous word. In the 'Word Cohesion' group, features representing the relationships between words and sentences within a paragraph increase the text's difficulty, demanding that readers link information within the same sentence and paragraph to form a complete data set.

While not significantly improving the model's performance like the three feature groups mentioned above, the' Syntax-Level Features' group still contributes to determining the sentence's difficulty through conjunction words. If the number of conjunction words is high, it creates multiple layers of relationships between subjects, a phenomenon present in the sentence. In contrast to the other feature groups, the 'Vietnamese-Specific Feature' group decreases the models' performance on all three datasets. This may be because the statistical features we used do not accurately reflect the nature of specific features present in Vietnamese. Sino-Vietnamese and borrowed words may indicate different semantic layers depending on usage, context, and the reader's existing knowledge. Therefore, determining the features of Sino-Vietnamese and borrowed words through a statistical approach may not be suitable.

Table 8 from the paper provides a comparative analysis of the accuracies achieved by different machine learning models across three datasets—Luong, OneStopEnglish, and RACE—with varying amounts of data (25%, 50%, and 75%). For the Luong dataset, the PhoBERT + MLP model shows a significant improvement in accuracy as the data size increases, while Random Forest and PhoBERT + Random Forest demonstrate remarkably high accuracy across all data sizes. In the case of OneStopEnglish, PhoBERT + MLP show increased accuracy with more data, but the performance is notably lower than on the Luong dataset, with PhoBERT + SVM even decreasing in accuracy as more data is provided. This could be explained that the OneStopEnglish dataset has only 567 samples, Extra Trees—a model that can capture complex patterns—might be overfitting to the training data at smaller data sizes. For the RACE dataset, the models exhibit a general trend of decreased accuracy a bit with increased data, with PhoBERT + Extra Trees showing the least variation. This may be due to the translation come with noise when increasing the size of data that can affect the model's ability to make accurate

predictions. These findings underscore the importance of considering both the nature of the dataset and the volume of data when selecting models for text readability tasks. It appears that no single model consistently outperforms others across all datasets and data sizes, highlighting the necessity for tailored approaches in readability assessment.

6 Limitation

While this study marks a significant advancement in the assessment of Vietnamese text readability, there are several limitations that must be acknowledged. Firstly, the reliance on translated datasets from English (OneStopEnglish and RACE) may not fully capture the intrinsic linguistic and cultural nuances of Vietnamese, potentially affecting the generalizability of the findings. Another limitation is the scope of the datasets used. The Vietnamese Text Readability Dataset (ViRead) is robust but may not represent all genres and styles of Vietnamese text. This could limit the model's applicability to diverse types of Vietnamese writings. Moreover, the machine learning models employed, despite their efficacy, might still have inherent biases and limitations in understanding complex language structures and idiomatic expressions. Finally, the current study focuses on lexical and syntactic features without deeply exploring pragmatic and discourse-level features, which are crucial for comprehensive readability assessment.

These limitations highlight areas for future research, suggesting the need for more diverse and culturally rich Vietnamese datasets, exploration of additional language models, and a broader consideration of linguistic features for a more nuanced understanding of text readability in Vietnamese.

References

[1] Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)

[2] Flesch, R.: A new readability yardstick. Journal of applied psychology **32**(3), 221 (1948)

[3] Lee, B.W., Jang, Y.S., Lee, J.: Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10669–10686. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.834 . https://aclanthology.org/2021.emnlp-main.834

[4] Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Human Language Technologies 2007: the Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 460–467 (2007)

[5] Heilman, M., Collins-Thompson, K., Eskenazi, M.: An analysis of statistical models and features for reading difficulty prediction. In: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, pp. 71–79 (2008)

[6] Si, L., Callan, J.: A statistical model for scientific readability. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 574–576 (2001)

[7] Vajjala, S., Meurers, D.: 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, pp. 163–173 (2012)

[8] Vajjala, S., Lučić, I.: Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 297–304 (2018)

[9] Xia, M., Kochmar, E., Briscoe, T.: Text readability assessment for second language learners. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 12–22 (2016)

[10] Nguyen, L.T., Henkin, A.B.: A second generation readability formula for vietnamese. Journal of Reading **29**(3), 219–225 (1985)

[11] Luong, A.-V., Nguyen, D., Dinh, D.: A new formula for vietnamese text readability assessment. In: 2018 10th International Conference on Knowledge and Systems Engineering (KSE), pp. 198–202 (2018). IEEE

[12] Doan, N.-T., Le, T.-A.-T., Luong, A.-V., Dinh, D.: Combining latent semantic analysis and pre-trained model for vietnamese text readability assessment: Combining statistical semantic embeddings and pre-trained model for vietnamese long-sequence readability assessment. In: Proceedings of the 4th International Conference on Information Technology and Computer Communications. ITCC '22, pp. 45–52. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3548636.3548643 . https://doi.org/10.1145/3548636.3548643

[13] Nguyen, D.Q., Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037–1042. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.92 . https://aclanthology.org/2020.findings-emnlp.92

[14] Luong, A.-V., Nguyen, D., Dinh, D.: Building a corpus for vietnamese text readability assessment in the literature domain. Universal Journal of Educational Research **8**(10), 4996–5004 (2020)

[15] Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785–794 (2017)

[16] Tran, C.D., Pham, N.H., Nguyen, A.-T., Hy, T.S., Vu, T.: Videberta: A powerful pre-trained language model for vietnamese. In: Findings of the Association for Computational Linguistics: EACL 2023, pp. 1041–1048 (2023)

[17] Tran, T.O., Le Hong, P., *et al.*: Improving sequence tagging for vietnamese text using transformer-based neural models. In: Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, pp. 13–20 (2020)

[18] Chall, J.S., Dale, E.: Readability revisited: The new dale-chall readability formula. (No Title) (1995)

[19] Fry, E.: A readability formula for short passages. Journal of Reading **33**(8), 594–597 (1990)

[20] Lennon, C., Burdick, H.: The lexile framework as an approach for reading measurement and success. electronic publication on www. lexile. com (2004)

[21] Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL-International Journal of Applied Linguistics **165**(2), 97–135 (2014)

[22] Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pp. 193–200 (2004)

[23] Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 523–530 (2005)

[24] Pilán, I., Volodina, E., Johansson, R.: Rule-based and machine learning approaches for second language sentence-level readability. In: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 174–184 (2014)

[25] Cha, M., Gwon, Y., Kung, H.: Language modeling by clustering with word embeddings for text readability assessment. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2003–2006 (2017)

[26] Jiang, Z., Gu, Q., Yin, Y., Chen, D.: Enriching word embeddings with domain knowledge for readability assessment. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 366–378 (2018)

[27] Azpiazu, I.M., Pera, M.S.: Multiattentive recurrent neural network architecture for multilingual readability assessment. Transactions of the Association for Computational Linguistics **7**, 421–436 (2019)

[28] Deutsch, T., Jasbi, M., Shieber, S.M.: Linguistic features for readability assessment.

In: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–17 (2020)

[29] Meng, C., Chen, M., Mao, J., Neville, J.: Readnet: A hierarchical transformer framework for web article readability analysis. In: Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42, pp. 33–49 (2020). Springer

[30] François, T., Fairon, C.: An "ai readability" formula for french as a foreign language. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 466–477 (2012)

[31] Dell'Orletta, F., Montemagni, S., Venturi, G.: Read–it: Assessing readability of italian texts with a view to text simplification. In: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pp. 73–83 (2011)

[32] Hancke, J., Vajjala, S., Meurers, D.: Readability classification for german using lexical, syntactic, and morphological features. In: Proceedings of COLING 2012, pp. 1063–1080 (2012)

[33] Falkenjack, J., Mühlenbock, K.H., Jönsson, A.: Features indicating readability in swedish text. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 27–40 (2013)

[34] Pilán, I., Vajjala, S., Volodina, E.: A readable read: Automatic assessment of language learning materials based on linguistic complexity. arXiv preprint arXiv:1603.08868 (2016)

[35] Islam, Z., Mehler, A., Rahman, R.: Text readability classification of textbooks of a low-resource language. In: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pp. 545–553 (2012)

[36] Chatzipanagiotidis, S., Giagkou, M., Meurers, D.: Broad linguistic complexity analysis for greek readability classification. In: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 48–58 (2021)

[37] Nguyen, L.T., Henkin, A.B.: A readability formula for vietnamese. Journal of Reading **26**(3), 243–251 (1982)

[38] Nguyễn, Đ.T.N., Lương, A.-V., Điền, Đ.: Affection of the part of speech elements in vietnamese text readability. Acta Linguistica Asiatica **9**(1), 105–118 (2019)

[39] Luong, A.-V., Nguyen, D., Dinh, D., Bui, T.: Assessing vietnamese text readability using multi-level linguistic features. International Journal of Advanced Computer Science and Applications **11**(8) (2020)

[40] Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese

natural language processing toolkit. In: Liu, Y., Paek, T., Patwardhan, M. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 56–60. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-5012 . https://aclanthology.org/N18-5012

[41] He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations (2020)