A study of Vietnamese readability assessing through semantic and statistical features

Hung Tuan Le, Long Truong To, Manh Trong Nguyen, Quyen Nguyen, Trong-Hop Do

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

A study of Vietnamese readability assessing through semantic and statistical features

Hung Tuan Le[1,3,♣], **Long Truong To**[1,3,♣], **Manh Trong Nguyen**[1,3,♣]
Quyen Nguyen[2,3,♢], **Trong-Hop Do**[1,3,♠]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]International University, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
{21521101, 21520250, 21520343}@gm.uit.edu.vn♣
ntquyen@hcmiu.edu.vn♢ hopdt@uit.edu.vn♠

Abstract

Determining the difficulty of a text involves assessing various textual features that may impact the reader's text comprehension, yet current research in Vietnamese has only focused on statistical features. This paper introduces a new approach that integrates statistical and semantic approaches to assessing text readability. Our research utilized three distinct datasets: the Vietnamese Text Readability Dataset (ViRead), OneStopEnglish, and RACE, with the latter two translated into Vietnamese. Advanced semantic analysis methods were employed for the semantic aspect using state-of-the-art language models such as PhoBERT, ViDeBERTa, and ViBERT. In addition, statistical methods were incorporated to extract syntactic and lexical features of the text. We conducted experiments using various machine learning models, including Support Vector Machine (SVM), Random Forest, and Extra Trees and evaluated their performance using accuracy and F1 score metrics. Our results indicate that a joint approach that combines semantic and statistical features significantly enhances the accuracy of readability classification compared to using each method in isolation. The current study emphasizes the importance of considering both statistical and semantic aspects for a more accurate assessment of text difficulty 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.

1 Introduction

Exchanging information and knowledge through texts has led to the emergence of measuring 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 lower reading speed, poorer comprehension, and reduced capacity to connect information within the text. In recent years, text difficulty has been evaluated through linguistically motivated features, 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. Two main approaches for determining text difficulties have been proposed, namely statistical approach and machine learning or deep learning. In the former 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. (Flesch, 1948; Kincaid et al., 1975), where these features are extracted and evaluated through correlation analysis with the difficulty of a set of texts. The second approach, namely machine or deep learning approach, involves using neural models to represent the semantics present in the text, allowing for the assessment of text difficulty (Heilman et al., 2007, 2008; Lee et al., 2021; Si and Callan, 2001).

Studies addressing the problem by applying advanced 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 (Vajjala and Meurers, 2012), OneStopEnglish (Vajjala and Lučić, 2018), and Cambridge (Xia et al., 2016). In Vietnam, pioneering research in this area, such as that of (Nguyen and Henkin, 1985; Luong et al., 2018), and more recently (Doan et al., 2022), has applied PhoBERT, which is a pre-trained language model (Nguyen and Tuan Nguyen, 2020) designed specifically for Vietnamese, to address the problem. However, these studies assess text difficulty of sentences in isolation while overlooking features that span over an extended discourse, such as discourse relations or entity cohesion across a series of sentences.

Given the gap in previous literature on Vietnamese text readability assessment, this study scru-

tinizes 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 (Luong et al., 2020a), RACE (Lai et al., 2017), and OneStopEnglish (Vajjala and Lučić, 2018). 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 (Nguyen and Tuan Nguyen, 2020), ViDeBERTa (Tran et al., 2023), and ViBERT (Tran et al., 2020). The joint approach combining statistical and semantic features are shown to 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 Works

This section provides an in-depth analysis of global body of research addressing the challenges of readability (see section 2.1), with a particular focus on the existing study conducted within the Vietnamese context (see section 2.2).

2.1 Textual Readability

Research on textual readability has increasingly captured of scholars within the natural language processing domain. This interest is particularly evident in foundational English-language studies, such as those pioneered by Flesch, which adopted a statistical lens to investigate the problem. These early investigation focused on evaluating text readability by quantifying linguistic features such as syllable per word ratio. Later, in 1975, the readability index by Kincaid et al. was published based on the features of Flesch. In Chall and Dale (1995), 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 (Fry, 1990; Lennon and Burdick, 2004), along with syntactic features such as the height of the parse tree (Chall and Dale, 1995). However, the statistical approach remains limited in its ability to capture deeper linguistic features that critically influence text readability, such as discourse relations, cohesion, and rhetorical structure (Collins-Thompson, 2014).

As language models have advanced and training data volumes have expanded, a new approach to the readability problem has emerged. This approach harnesses the language representation capabilities of these models to extract deeper linguistic features while utilizing the classification power of probabilistic and 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 (Schwarm and Ostendorf, 2005; Heilman et al., 2007, 2008; Pilán et al., 2014). Since the rise of deep learning models, particularly with the advent of pre-trained language models utilizing transformer architecture, which have achieved state-of-the-art results across various semantic tasks, the performance on the readability problem has notably improved. This enhancement is due not only to the advanced feature extraction capabilities of these models (Cha et al., 2017; Jiang et al., 2018; Azpiazu and Pera, 2019) but also to their integration with externally collected statistical features (Deutsch et al., 2020; Meng et al., 2020; Lee et al., 2021).

Beyond English, research has also expanded to other languages, building upon the established foundation of English-language studies, with notable developments in languages such as French (François and Fairon, 2012), Italian (Dell'Orletta et al., 2011), German (Hancke et al., 2012), Swedish (Falkenjack et al., 2013; Pilán et al., 2016), Bangla (Islam et al., 2012), and Greek (Chatzipanagiotidis et al., 2021).

2.2 Vietnamese Readability

Research on the readability problem remains limited, primarily due to the scarcity of high-quality datasets. This issue is evident in studies ranging from (Nguyen and Henkin, 1985, 1982) to (Luong et al., 2020a, 2018; Nguyễn et al., 2019), where dataset sizes have been notably small, often comprising fewer than 2,000 samples. Furthermore, the dominant approach to addressing the readability

problem has centered on feature extraction through statistical analysis. This includes metrics such as the number of syllables or words, the height and width of parse trees, and the count of clauses (Luong et al., 2020b). Recently, Doan et al. adopted a novel approach to the problem by extracting features using PhoBERT (Nguyen and Tuan Nguyen, 2020). However, this research has yet to be made accessible to the broader community.

3 Current Study

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 three datasets described in Table 1, namely OneStopEnglish (Vajjala and Lučić, 2018), RACE (Lai et al., 2017), and the Vietnamese Text Readability Dataset (Luong et al., 2020a).

The Vietnamese Text Readability Dataset (ViRead) (Luong et al., 2020a) is constructed from Vietnamese college-level textbooks, stories, and literature websites. After extracting text from these sources using OCR, a team of twenty Vietnamese literature teachers from middle schools, high schools, and colleges labels the sentences. The labels are categorized into four levels: 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 (Vajjala and Lučić, 2018) and RACE (Lai et al., 2017). The OneStopEnglish dataset is extracted from onestopenglish[1], an English language learning resources website run by MacMillan Education. The content has been rewritten into three versions from *The Guardian* newspaper, each labeled as advanced (Adv), intermediate (Int), and elementary (Ele). The RACE dataset, a large-scale reading comprehension benchmark, is derived from English exams administered to Chinese middle and high school students and includes 28,000 passages. For the readability task, RACE is divided into junior and senior levels.

We translated the two English datasets, OneStopEnglish and RACE, into Vietnamese using Google Translate[2]. Subsequently, we partitioned these datasets into smaller components for the experimentation process. Given the limited size of the OneStopEnglish and ViRead datasets, each containing fewer than 2,000 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.

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 employing statistical methods to extract features from the dataset, whereas the semantic approach leverages machine learning models, ranging from basic to advanced deep learning techniques, to derive semantic features. Additionally, we conduct experiments that integrate features from both statistical and semantic approaches to examine their correlation and impact on the results (see section 3.2.3).

3.2.1 Statistical approach

Luong et al. performed experiments to evaluate the impact of various features on text readability using a statistical approach, specifically on the Vietnamese readability dataset (Luong et al., 2020a). The features examined included part-of-speech features (such as the ratio of POS-tagged words and the proportion of common nouns to distinct words), syntax-level features (including average parse tree depths), and Vietnamese-specific features (like the ratio of borrowed words and Sino-Vietnamese words). We selected features that exhibited a high correlation with text difficulty, as detailed in Table 2.

Additionally, we introduced two new features related to word cohesion, represented through dependency trees, to investigate how the relationships between words within a sentence impact text difficulty (see table 2). To extract these two features, we utilized VnCoreNLP (Vu et al., 2018) for sentence segmentation and dependency representation. The statistical features will be classified using three machine learning models: Support Vector Machine (SVM), Random Forest, and Extra Trees.

The statistical features on the three datasets ViRead, OneStopEnglish, and RACE are summarized in Table 3. As noted, in translated datasets such as OneStopEnglish and RACE, some standard text features remain consistent, such as 'Average

[1]https://onestopenglish.com/

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

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

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
Syntax-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

word length in characters' and 'Distinct parallel conjunctions/distinct words.' For the 'Ratio of long sentences' feature, we define sentences with more than 20 syllables, based on research from the American Press Institute. However, features specific to Vietnamese, such as the 'Ratio of borrowed words' and the 'Ratio of distinct Sino-Vietnamese words,' vary. This variation is attributed to translation nuances and unique characteristics of Vietnamese texts. These differences significantly impact 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 (Nguyen and Tuan Nguyen, 2020), ViDeBERTa (Tran et al., 2023), and ViBERT (Tran et al., 2020). These models are instrumental in extracting deep semantic features from the Vietnamese texts, which are crucial for our classification task.

PhoBERT (Nguyen and Tuan Nguyen, 2020) emerges as a paragon, trained extensively on a corpus comprising 20GB of Vietnamese Wikipedia and news texts. It boasts 135 million parameters in its base iteration and an augmented 370 million parameters for the large variant. In its most recent iteration, PhoBERT$_{base} - V2$, the model has been refined on a formidable 120GB of Vietnamese texts derived from the OSCAR-2301 dataset[3].

ViDeBERTa (Tran et al., 2023) is a model with the architecture of DeBERTa (He et al., 2020) 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 (Tran et al., 2020) 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 features extracted from pre-trained language models will be classified using a range of machine

[3]https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

[4]https://huggingface.co/datasets/cc100

[5]https://huggingface.co/Fsoft-AIC/videberta-base

Feature	ViRead	OneStopEnglish	RACE
Number of words	40 - 23104	263 - 1417	13 - 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

learning models, including 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 explore the synergy between statistical and semantic approaches by conducting experiments that combine features from both methods. The goal of these experiments is to understand the complementary nature of these approaches and how their integration can 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:

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

4 Experiment Result

4.1 Statistical Result

The results presented in Table 4 reveal the Extra Tree model performs exceptionally well on both the OneStopEnglish and RACE datasets. On the OneStopEnglish dataset, Extra Tree surpasses the other models, SVM and Random Forest, by 0.8% in F_1-score compared to the second-best model (Random Forest) and by 2.92% compared to SVM. In the RACE dataset, Extra Tree continues to be the top performer. 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 variation in performance between Extra Tree and the other models across the two datasets is likely due to the substantial difference in dataset sizes, with OneStopEnglish comprising only 567 samples, while RACE contains 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 observed with the ViRead dataset can be attributed to the fact that the RACE and OneStopEnglish datasets are translations from English to Vietnamese. This translation process results in fewer features that are unique to Vietnamese compared to ViRead, which is derived from Vietnamese-language textbooks and thus retains more distinctive linguistic features inherent to Vietnamese.

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

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

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

classification models such as Random Forest and Extra Tree, features 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

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

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 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

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 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 Conclusion

In this paper, we propose a novel approach to the Vietnamese readability task by incorporating semantic features alongside traditional statistical features, leading to promising results on readability datasets. Additionally, we examine the impact of combining both feature types to enhance the performance of existing models. Our research has the potential to support the development of readability assessment systems for elementary-level writing. Using our model, educators can gain clear insights into the strengths and limitations of young students' essays, thereby aiding the learning and writing process in early education. Beyond this, our research shows promise in developing systems that suggest quality improvements for essays or even detect essays generated automatically by large language models.

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

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Miriam Cha, Youngjune Gwon, and HT Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006.

Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited: The new dale-chall readability formula. *(No Title)*.

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.

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

Kevyn Collins-Thompson and James P Callan. 2004. 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*, pages 193–200.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. 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*, pages 73–83.

Tovly Deutsch, Masoud Jasbi, and Stuart M 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.

Nam-Thuan Doan, Thi-Anh-Thi Le, An-Vinh Luong, and Dien Dinh. 2022. 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, page 45–52, New York, NY, USA. Association for Computing Machinery.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François and Cédrick Fairon. 2012. 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*, pages 466–477.

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

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. 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*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. 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*, pages 71–79.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. 2018. Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colleen Lennon and Hal Burdick. 2004. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*.

An-Vinh Luong, Diep Nguyen, and Dien Dinh. 2018. A new formula for vietnamese text readability assessment. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 198–202. IEEE.

An-Vinh Luong, Diep Nguyen, and Dien Dinh. 2020a. Building a corpus for vietnamese text readability assessment in the literature domain. *Universal Journal of Educational Research*, 8(10):4996–5004.

An-Vinh Luong, Diep Nguyen, Dien Dinh, and Thuy Bui. 2020b. Assessing vietnamese text readability using multi-level linguistic features. *International Journal of Advanced Computer Science and Applications*, 11(8).

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. 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*, pages 33–49. Springer.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Liem T Nguyen and Alan B Henkin. 1985. A second generation readability formula for vietnamese. *Journal of Reading*, 29(3):219–225.

Liem Thanh Nguyen and Alan B Henkin. 1982. A readability formula for vietnamese. *Journal of Reading*, 26(3):243–251.

Điệp Thi Nhu Nguyễn, An-Vinh Lương, and ĐINH Điền. 2019. Affection of the part of speech elements in vietnamese text readability. *Acta Linguistica Asiatica*, 9(1):105–118.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. 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*, pages 174–184.

Sarah E Schwarm and Mari Ostendorf. 2005. 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)*, pages 523–530.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Cong Dao Tran, Nhut Huy Pham, Anh-Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. Videberta: A powerful pre-trained language model for vietnamese. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1041–1048.

Thi Oanh Tran, Phuong Le Hong, et al. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia conference on language, information and computation*, pages 13–20.

Sowmya Vajjala and Ivana Lučić. 2018. 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*, pages 297–304.

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.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

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.

A Analysis of different features on the performance

In Table 7, we present the impact of different feature groups on the performance of models that combine both embedding and statistical features. These experiments were conducted using the best-performing models from each dataset. The results demonstrate that the "Raw Feature" group has the most significant effect on model performance, followed by the "POS Feature" and "Word Cohesion" groups. In contrast, "Syntax-Level" and "Vietnamese-Specific" features contribute less to performance improvement, with Vietnamese-specific features sometimes leading to decreased performance compared to raw features.

B Analysis of performance based on the data size

Table 8 presents a comparison of model accuracies across datasets with varying data sizes (25%, 50%, and 75%). The results demonstrate how accuracy trends vary depending on the dataset and the model used. While PhoBERT-based models like PhoBERT + MLP show consistent improvement with larger data sizes in most cases, others

Dataset	Model	Raw		POS		Syntax-Level		Viet-Spec		Word Coh.	
		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 + RF	**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 + ET	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

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

like Random Forest exhibit stable high accuracy across all data sizes.