VLUE: A New Benchmark and Multi-task Knowledge Transfer Learning for Vietnamese Natural Language Understanding

Anonymous EMNLP submission

Abstract

The success of Natural Language Understanding (NLU) benchmarks in various languages, such as GLUE (Wang et al., 2018) for English, CLUE (Xu et al., 2020) for Chinese, KLUE (Park et al.) for Korean, and IndoNLU (Wilie et al., 2020) for Indonesian, has facilitated the evaluation of new NLU models across a wide range of tasks. To establish a standardized set of benchmarks for Vietnamese NLU, we introduce the first Vietnamese Language Understanding Evaluation (**VLUE**) benchmark. The VLUE benchmark encompasses five datasets covering different NLU tasks, including text classification, span extraction, and natural language understanding. To provide an insightful overview of the current state of Vietnamese NLU, we then evaluate seven state-of-the-art pre-trained models, including both multilingual and Vietnamese monolingual models, on our proposed VLUE benchmark. Furthermore, we present **CafeBERT**, a new state-of-the-art pre-trained model that achieves superior results across all tasks in the VLUE benchmark. Our model combines the proficiency of a multilingual pre-trained model with Vietnamese linguistic knowledge. CafeBERT is developed based on the XLM-RoBERTa model, with an additional pretraining step utilizing a significant amount of Vietnamese textual data to enhance its adaptation to the Vietnamese language. For the purpose of future research, CafeBERT is made publicly available[1] for research purposes.

1 Introduction

Recently, the Vietnamese Natural Language Processing (NLP) research community has achieved remarkable advancements in the development of pre-trained language models for the Vietnamese language (Nguyen and Tuan Nguyen, 2020; Tran et al., 2022, 2023). The integration of these state-of-the-art models, coupled with the progress made in establishing high-quality benchmarks, has paved the way for a diverse array of applications within Vietnam. Notably, these advancements have greatly enhanced capabilities in areas of Machine Reading Comprehension (Van Kiet et al., 2022; Van Nguyen et al., 2021).

Unfortunately, despite the recent progress in developing large language models for Vietnamese, the research community of Vietnamese NLP lacks a common ground for evaluating the performance of these models. This lack of standard evaluation metrics and benchmarks makes it difficult to identify the strengths and weaknesses of different approaches in pre-training new models in Vietnamese and the overall progress of Vietnamese natural language understanding (NLU). As a result, it is crucial for the community to establish a shared set of evaluation metrics and benchmarks that can be used to assess newly proposed language models. Inspired by benchmarks evaluating Natural Language Understanding in other languages (Wang et al., 2018, 2019; Xu et al., 2020; Wilie et al., 2020; Park et al.), in this paper, we propose VLUE (Vietnamese Language Understanding Evaluation) as a shared set of evaluation metrics and benchmarks for pre-trained models in Vietnamese. To the best of our knowledge, our proposed benchmark is the first benchmark for evaluating Vietnamese NLU models. We believe that this benchmark will serve as a valuable resource for researchers and practitioners working in the field of Vietnamese NLU, and will help drive further advancements in this area.

To facilitate the development of new large language models in Vietnamese, we, in this work, introduce Vietnamese Language Understanding Evaluation (**VLUE**), a comprehensive language understanding framework that includes five diverse tasks. The tasks include a wide range of applications (Question Answering, Hate Speech Detection, Part-of-Speech, Emotion Recognition, and Natural

[1] The link will be provided upon acceptance.

Language Inference), types of input (single sentences, pair of sentences, sequence of sentences) and objectives of tasks (extracted span, sentence classification, sequence labeling). With its diverse set of benchmarks, VLUE establishes a standardized evaluation framework, enabling comprehensive comparisons and evaluations of different models in the context of Vietnamese.

Within this paper, we commence by introducing our novel VLUE benchmark, designed to evaluate the language prowess of various models. We conduct a comprehensive analysis of seven models, encompassing four multilingual models as well as three monolingual models. Additionally, we present the introduction of a newly developed pre-trained model, referred to as **CafeBERT**. This model is constructed by leveraging the large-scale XLM-RoBERTa model and further fine-tuning it on an extensive Vietnamese corpus, thereby enhancing its proficiency in the Vietnamese language and elevating its overall performance. Through in-depth evaluation, we demonstrate that **CafeBERT** achieves state-of-the-art performance across all four tasks presented in our VLUE benchmark.

In this paper, we make the following contributions:

1. Our paper introduces a high-quality Vietnamese natural language understanding benchmark that covers a variety of tasks: Part-of-speech tagging, machine reading comprehension, natural language inference and hate speech spans detection, at different levels of difficulty, in different sizes and domains. This benchmark serves as a common ground for assessing the overall proficiency of language models in the Vietnamese language.

2. We propose an enhanced version of XLM-RoBERTa large that is specifically optimized for Vietnamese. Through comprehensive testing on the VLUE benchmark, we show that our model substantially outperforms existing models. We publicly release our models under the name **CafeBERT** which can serve as a strong baseline for future Vietnamese computational linguistics research and applications.

The rest of this paper is structured as follows. Section 2 reviews existing NLU benchmarks and pre-trained language models. Section 3 introduces the NLU benchmark for Vietnamese. In particular, we present experiments and benchmark result in Section 4. Then Section 5 presents a new pre-trained language model called CafeBERT. Finally, Section 6 presents conclusions and future work.

2 Related Work

In this paper, we review data benchmark and pre-trained language models related to our work.

2.1 Benchmarks

This work is directly inspired by GLUE benchmark (Wang et al., 2018) which is a multi-task benchmark for natural language understanding (NLU) in the English language. It consists of nine tasks: single-sentence classification, similarity and paraphrase tasks, and Inference Tasks. Later, recognizing that performance of SOTA models on the benchmark has recently surpassed the level of non-expert humans, suggesting limited headroom for further research, Wang et al. (2019) propose SuperGLUE which is GLUE's harder counterpart. SuperGLUE covers question answering, NLI, co-reference resolution, and word sense disambiguation tasks.

Following the idea of GLUE and SuperGLUE, different NLU benchmarks are also introduced in other languages such as CLUE (Xu et al., 2020) in Chinese, FLUE (Le et al., 2020) in French, IndoNLU (Wilie et al., 2020) in Indonesian. Besides, in the multilingual setting, we also have XGLUE (Liang et al., 2020) for evaluating Cross-lingual Pre-training, Understanding and Generation.

2.2 Pretrained Language Models

Pre-trained language models have revolutionized the field of natural language processing (NLP) by providing a powerful foundation for various language-related tasks. These models are typically designed based on the architecture of the Transformers model (Vaswani et al., 2017), which has proven to be highly effective in capturing intricate patterns and dependencies in textual data by utilizing attention mechanisms.

The concept of pre-training involves training models using large amounts of text data in semi-supervised tasks. During pre-training, the models learn to predict missing words (Masked Language Model) or determine the coherence between pairs of sentences (Next Sentence Prediction) (Devlin et al., 2019). By learning from diverse and vast text corpora, these models acquire a rich understanding of language, including grammar, semantics, and

contextual cues.

Following the groundbreaking success of BERT (Devlin et al., 2019), a wave of enhanced variations has emerged, each pushing the boundaries of pre-trained language models. Noteworthy among these advancements are RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), SpanBERT (Joshi et al., 2020), and DeBERTa (He et al., 2021) are developed. Additionally, several BERT variants have been developed for multilingual applications in over 100 languages, such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020a).

Following the wave of pre-training in English, researchers worldwide have embarked on pre-training monolingual language models in diverse languages. This linguistic expansion has resulted in the development of notable models like CamemBERT (Chan et al., 2020) in French, GELECTRA (Martin et al., 2020) in German, and BERT and its variations (Cui et al., 2021) in Chinese.

3 VLUE Benchmark

3.1 Overview

VLUE is a collection of five language understanding tasks in Vietnamese. The goal of VLUE is to provide a set of high-quality benchmarks to assess the Vietnamese language understanding of newly proposed models. The selected tasks are guaranteed through many criteria to make the most accurate assessment. VLUE covers a wide variety of tasks with variations in the size of the dataset, the size of the input text, and the comprehension requirements of each task. The datasets should be easy to implement for evaluation so that users can focus on developing models. The selected tasks are challenging for the model but must be solvable. Table 1 presents the overview of the datasets and tasks in VLUE. Data samples for each task are shown in Table 5. We describe each dataset and task as follow.

3.2 Tasks

UIT-ViQuAD 2.0 The Vietnamese Question Answering Dataset 2.0 (Van Kiet et al., 2022) is an updated version of the UIT-ViQuAD 1.0 dataset (Nguyen et al., 2020). UIT-ViQuAD 2.0 is published for the machine reading comprehension shared-task at the Eighth Workshop on Vietnamese Language and Speech Processing (VLSP 2021). This dataset includes $5,173$ paragraphs extracted from 176 articles on the Wikipedia data domain. The hired human annotators then annotate $24,489$ answerable questions and $11,501$ unanswerable questions. The task proposed by this dataset is to extract the answer for a question given a corresponding context. The answer can be empty when models encounter unanswerable questions. Exact Match (EM) and F1-score are used to evaluate the performance of the model.

ViNLI The Vietnamese Natural Language Inference dataset (Huynh et al., 2022) is the first Vietnamese high-quality and large-scale dataset created for the open-domain natural language inference task. The dataset consists of more than $30,000$ human-annotated premise-hypothesis sentence pairs with 13 topics from more than 800 online news articles. The goal of the problem is to predict the relationship of pairs of sentences with the set of relationships that include entailment, neutral, contradiction, and other. Following the original work of ViNLI, we use F1-score and Accuracy as the metrics for the evaluation process.

VSMEC The standard Vietnamese Social Media Emotion Corpus (Ho et al., 2020), or UIT-VSMEC (VSMEC), is the task of classifying the emotion of Vietnamese comments on social networks. The dataset includes $6,927$ manually labeled social media comments. It is a multi-label classification problem with seven emotion labels: anger, disgust, enjoyment, fear, sadness, surprise, and other. Enjoyment label has the most significant rate with about 28%, and surprise is the lowest with less than 5%. Following (Nguyen et al., 2022), the F1-macro is used as a metric to evaluate VSMEC.

ViHOS The Vietnamese Hate and Offensive Span dataset (Hoang et al., 2023) consists of $26,467$ spans on $11,056$ comments (including clean, hate, and offensive comments). The dataset is annotated by humans through three labeling phases. The goal of this task is to extract hate and offensive spans from comments. The dataset is a challenge as about 51% of comments have no span extracted and about 27% of comments have more than one extracted hate speech spans. F1-score is the metric used in this dataset to evaluate the performance of the model.

NIIVTB POS NIIVTB (Nguyen et al., 2016, 2018b) is a constituent treebank in Vietnamese annotated with three layers: word segmentation, part-of-speech (POS), and bracketing. In the VLUE benchmark, we use the POS task in NIIVTB, so we call NIIVTB POS. This treebank has two sub-

Dataset	Train	Dev	Test	Domain	Task	Metric
UIT-ViQuAD	28,457	3,821	3,712	Wikipedia	Machine reading comprehension	EM / F1
ViNLI	24,376	3,009	2,991	Online news	Natural language inference	Acc / F1
VSMEC	5,548	686	693	Social networks	Emotion recognition	F1
ViHOS	8,974	1,112	1,128	Social networks	Hate speech spans detection	F1
NIIVTB POS	18,588	1,000	1,000	Online news	Part-of-speech tagging	F1

Table 1: Statistics of the VLUE datasets and tasks. The version of UIT-ViQuAD is 2.0. ViNLI has four classes.

sets, NIIVTB-1 and NIIVTB-2, with more than $10,000$ sentences each crawled from two sources: the first set is VLSP[2] raw data from Youth[3] (Tuổi Trẻ) online newspaper with the topic are social and political topics, the second set is collected from Thanhnien[4] online newspaper with 14 different topics. NIIVTB has $20,588$ sentences divided into three sets of train, dev, and test with a ratio of roughly $8:1:1$. We use F1 as the metric for evaluating the POS task of NIIVTB.

4 Experiments and Benchmark Result

4.1 Experiment settings

Baselines To provide an insightful overview of the current progress of Vietnamese NLU, we implement state-of-the-art models in Vietnamese NLU using the library *Transformers* provided by Huggingface[5]. For the text classification task, we encode the input sentence and then pass the encoded output through a classifier. Similar to text classification tasks, for NLI tasks, we encode the input sentence pair with a separator token and then pass the output through a classifier. For span extraction tasks, we use two fully connected layers after encoding the input to predict the start and end position of the segment to be extracted.

All of our experiments are performed on a single machine with an NVIDIA A100 GPU with 40GB of RAM on a Google Colaboratory environment[6]. We use TensorFlow 2.11.0 (Abadi et al., 2016) and PyTorch 1.12.0 (Paszke et al., 2019) to support the research process.

Models We use the public available pre-trained models that support Vietnamese below to evaluate models on VLUE benchmark. The details of each model are shown in Table 2.

- **mBERT** (Devlin et al., 2019): We use base version model with 12 layers and hidden size of 768. The model has been trained with big data corpus covering 104 languages including Vietnamese.

- **WikiBERT** (Pyysalo et al., 2021): WikiBERT for Vietnamese belongs to a group of 42 WikiBERT models that support 42 different languages. Vietnamese WikiBERT is built using the BERT architecture and trained using data from two sources: Wikipedia (172M tokens) and the Vietnamese Treebank dataset ($20,285$ tokens).

- **DistilBERT** (Sanh et al., 2019): DistilBERT was introduced as a smaller, lighter, and faster version of the previous BERT model but retained 97% of its language comprehension. Multilingual DistilBERT is trained in 104 languages with a hidden size of 768 and 6 layers.

- **PhoBERT** (Nguyen and Tuan Nguyen, 2020): PhoBERT is the state-of-the-art monolingual model in Vietnamese. The model is trained based on the RoBERTa model with a dataset including Vietnamese Wikipedia and news articles. PhoBERT has two versions, including PhoBERT$_{base}$ and PhoBERT$_{large}$.

- **XLM-RoBERTa** (Conneau et al., 2020b): XLM-RoBERTa is a large-scale pre-trained multilingual model. This model was trained on a Transformers-based masked language task using two terabytes of CommonCrawl data across more than a hundred languages. The model has two versions, XLM-RoBERTa$_{base}$ and XLM-RoBERTa$_{large}$.

These models currently achieve state-of-the-art performance on most Vietnamese language processing benchmarks. Among the models above, the multilingual model XLMR$_{large}$ and monolingual model PhoBERT$_{large}$ are the two most important models in Vietnamese NLP at the time of

[2]https://vlsp.hpda.vn/demo/
[3]https://tuoitre.vn/
[4]https://thanhnien.vn/
[5]https://huggingface.co/
[6]https://colab.research.google.com/

Model	#Params	#Layers	#Heads	Hidden Size	Vocab Size	Language Type	Data Pre-train Source
wikiBERT	-	12	12	768	20101	monolingual	Wikipedia
PhoBERT$_{base}$	135M	12	12	768	64001	monolingual	Wikipedia, News
PhoBERT$_{large}$	370M	24	16	1024	64001	monolingual	Wikipedia, News
mBERT	179M	12	12	768	119547	multilingual	Wikipedia
DistilBERT	134M	6	12	768	119547	multilingual	Wikipedia
XLM-Roberta$_{base}$	270M	12	8	768	250002	multilingual	CommonCrawl
XLM-Roberta$_{large}$	550M	24	16	1024	250002	multilingual	CommonCrawl
CafeBERT	550M	24	16	1024	250002	multilingual	Wikipedia, News

Table 2: The details of baseline models used in VLUE benchmark.

this writing and are expected to achieve impressive performance on VLUE benchmark tasks.

4.2 Result Benchmark

Table 3 presents the results of all experimented models on the VLUE tasks. We observed that the larger the model, the higher the performance, typically the XLM-Roberta$_{large}$ and PhoBERT$_{large}$ models with the most significant number of parameters have outstanding performance on all tasks. XLM-RoBERTa$_{large}$ is the model with the best performance on 4 over 5 VLUE tasks including UIT-ViQuAD, ViNLI, ViHOS, and NIIVTB POS. This results agree with multiple previous work as XLM-Roberta$_{large}$ also achieves SOTA results other Vietnamese tasks other than the VLUE benchmark (Do et al., 2021; Van Nguyen et al., 2023; Tran et al., 2021). PhoBERT$_{large}$ is the model with the best performance on VSMEC tasks with F1-score achieved is 65.44%. Especially for the NIIVTB POS task, the pre-trained multilingual models have higher performance than the pre-trained monolingual models. XLM-Roberta$_{large}$ has the highest performance on NIIVTB POS, with an 83.62% F1-score.

According to the results, models pre-trained on multilingual data perform better than monolingual pre-trained models. The multilingual model XLM-Roberta performed better than the best pre-trained model - PhoBERT, in 4 tasks of the VLUE benchmark. For the base version of the two models above, PhoBERT is stronger than XLM-Roberta with a ratio of 3: 2: PhoBERT on UIT-ViQuAD 2.0, ViNLI, and ViHOS; XLM-RoBERTa on VSMEC and NIIVTB POS. The above result is because the number of attention heads of XLM-Roberta is eight which is less than twelve of PhoBERT.

We then compare WikiBERT (monolingual pre-trained model) and mBERT (multilingual pre-trained model), the two models with the same number of attention heads and the number of layers (transformers block). We observe that mBERT outperforms WikiBERT on three tasks (UIT-ViQuAD 2.0, ViNLI, NIIVTB POS), similar to results from work in other languages (Pikuliak et al., 2022; Armengol-Estapé et al., 2022).

The monolingual pre-training models perform better than the multilingual pre-training models in the social network domain (Quoc Tran et al., 2023; Nguyen et al., 2022). In the VLUE benchmark, there are two models with a social network domain, VSMEC, and ViHOS. For VSMEC, the PhoBERT large model achieve the SOTA results. With the ViHOS dataset, the XLM-RoBERTa model achieve the best performance. However, the difference in results between XLM-RoBERTa and PhoBERT is minor (only 0.54%) compared to the difference between the two models in other tasks ranging from 3% to 6%. This results suggest that training NLU models with monolingual textual data is necessary for tasks whose domain is social networks (Wilie et al., 2020; Müller et al., 2020). On the other hand, models trained with multilingual data can comprehend multiple languages and tackle tasks that involve corpora with a significant presence of foreign words (non-Vietnamese), such as news articles and Wikipedia.

5 CafeBERT

The results from our analysis on current progress of Vietnamese NLU show that the XLM-RoBERTa$_{large}$ achieves the best performance on most tasks of VLUE. However, PhoBERT also show a comparable performance on tasks with corpus from social networks, such as VSMEC and ViHOS. This observation drives us to a hypothesis that further adapting multilingual model XLM-RoBERTa$_{large}$ into Vietnamese can help improve

Models	UIT-ViQuAD 2.0		ViNLI		VSMEC	ViHOS	NIIVTB POS
	EM	F1	Accuracy	F1	F1	F1	F1
Human	75.50	82.85	95.78	95.79	-	-	-
wikiBERT [✦]	42.16	52.62	71.18		57.64	77.05	75.52
PhoBERT$_{base}$ [✦]	51.00	64.29	78.00	78.05	59.91	75.69	77.60
PhoBERT$_{large}$ [✦]	57.27	70.88	80.67	80.69	65.44	77.16	79.36
mBERT [✧]	52.34	63.71	73.45	73.62	54.59	76.22	81.34
DistilBERT [✧]	53.83	35.78	44.39	66.77	53.83	75.72	80.05
XLM-Roberta$_{base}$ [✧]	50.49	59.23	76.83	77.01	61.89	74.67	81.76
XLM-Roberta$_{large}$ [✧]	64.71	75.36	85.99	86.10	62.24	77.70	83.62
CafeBERT	**65.25**	**76.36**	**86.11**	**86.16**	**66.12**	**78.56**	**84.04**

Table 3: Baseline performance on the VLUE benchmark. For the UIT-ViQuAD dataset, we report EM (the rate of match between the gold and predicted answers) and F1. For the the ViNLI dataset, we report Accuracy and F1. For the ViHOS dataset, we report F1. For the NIIVTB POS dataset, we report F1. *Avg* is the average of all tasks. The best results for each task are in **bold** text. [✦] and [✧] are monolingual model and multilingual model, respectively.

its performance on VLUE. We then propose a new model that is expected to combine the existing knowledge from XLM-RoBERTa and the newly trained knowledge from Vietnamese corpus. We continue pre-training XLM-RoBERTa with a Vietnamese dataset similar to the data used to train the PhoBERT model. We refer to our proposed model as CafeBERT.

5.1 Dataset and Training New Language Model

In this section, we describes the dataset, architecture, and training setting that we used to develop the new pre-training model.

Pre-training data: We use a corpus of 18GB of textual data as the pre-training dataset. The dataset has two corpora: 1GB of text from the Vietnamese Wikipedia and 17GB of text which is de-duplicated and preprocessed data from a 27.5GB corpus of text sourced from online Vietnamese news articles[7]. Our dataset contains about 180 million sentences and more than 2.8 billion word tokens.

Architecture: Our model is built upon the XLM-Roberta model (Conneau et al., 2020b) by continue pre-training it on the large Vietnamese text corpus. The training process uses the objective of the mask language model (MLM) task. Our model has a hidden state of 1024, 24 layers, and 16 attention heads.

Fine-tuning: We create the CafeBERT pre-training model by fine-tuning the XLM-Roberta model with the transformers library[8]. The optimizer for training is Adam (Kingma and Ba, 2014) with weight decay (Loshchilov and Hutter, 2019). We fine-tuned the model on an A100 40GB GPU with a peak learning rate of 2e-5. For the MLM task, we do masking for 15% of the words of the data.

[7]https://github.com/binhvq/news-corpus

[8]https://github.com/huggingface/transformers

5.2 Results of CafeBERT

5.2.1 Results of CafeBERT on VLUE

Table 3 shows that our new pre-trained model achieves best performance on all the tasks of the VLUE benchmark. On UIT-ViQuAD 2.0 dataset, CafeBERT has the best improvement in F1-score with a 1% increase on the test set. On the other hand, this model has a minor performance increase with 0.06% F1-score and 0.12% accuracy on the test set of ViNLI. On the VSMEC dataset, our pre-trained model CafeBERT outperforms PhoBERT$_{large}$ by 0.68% F1-score and 3.88% F1-score over XLM-Roberta$_{large}$. On ViHOS and NIIVTB POS datasets, CafeBERT achieves the new SOTA results with F1-scores on the test set of 78.56% (+0.86%) and 84.04% (+0.42%), respectively. Besides, CafeBERT also performs well on all corpus domains in VLUE, including Wikipedia, news, and social networks. So our model sets a new SOTA performance on the VLUE benchmark and establishes a strong baseline for future proposed Vietnamese NLU model.

5.2.2 Results of CafeBERT on other tasks

In addition to the tasks in VLUE, we implement the CafeBERT model on other tasks in Vietnamese including: ViNewsQA, UIT-ViFSD, and UIT-VSFC. In which:

- **ViNewsQA** (Nguyen et al., 2021) is an ma-

Models	ViNewsQA		UIT-ViSFD	UIT-VSFC			
				Sentiment Classification		Topic Classification	
	EM	F1	F1	Accuracy	F1	Accuracy	F1
wikiBERT	62.30	82.85	71.46	-	-	-	-
PhoBERT$_{large}$	70.98	88.89	77.52	93.43	82.81	88.22	78.08
mBERT	63.81	83.19	70.27	91.88	78.67	87.93	77.28
distilBERT	-	-	70.97	-	-	-	-
XLM-Roberta$_{large}$	71.49	89.44	82.51	94.13	83.70	88.57	79.20
CafeBERT	77.53	91.39	83.13	94.16	84.29	89.07	79.82

Table 4: Performance of models on tasks outside VLUE. We evaluate the results on the test data set.

chine reading comprehension task on the health domain. The dataset contains 22,057 question-answer pairs extracted from health news.

- **UIT-ViFSD** (Phan et al., 2021) is the customer comments classification on e-commerce platforms. The data set includes 11,122 comments about phones classified into three sentiments: positive, negative, and neutral.
- **UIT-VSFC** (Nguyen et al., 2018a) is a dataset including 16,000 student feedback sentences. Sentences are human-annotated with two tasks: sentiment-based classification and topic-based classification.

Table 4 shows our experimental results on the three datasets described above with several pre-trained models that support Vietnamese. On all three tasks, the CafeBERT model has better results than other models. In tasks C and D, the CafeBERT model has higher performance than the model with the second best results (XLM-Roberta$_{large}$) by just under 1% in evaluation metrics. The CafeBERT model shows the highest superiority in the ViNewsQA task with F1 and accuracy 1.95% and 6.04% higher, respectively, when compared to the XLM-Roberta$_{large}$ model. The CafeBERT model is enhanced by training on corpus text mainly in news domains similar to ViNewsQA's data source, so the CafeBERT model shows its best power on this task.

6 Conclusion and Future Works

We proposed **VLUE** - the first Vietnamese language understanding evaluation benchmark. VLUE is used to evaluate pre-trained models in Vietnamese with various tasks such as reading comprehension, text classification, natural language inference, hate speech detection, and part-of-speech tagging. We also publicize a pre-trained model, **CafeBERT**, which is trained based on the XLM-Roberta model with a vast Vietnamese text dataset. We show that CafeBERT achieves SOTA performance on all VLUE benchmark tasks and all VLUE domains, such as social networks, Wikipedia, and news.

We expect VLUE to be widely used to evaluate Vietnamese-supported pre-trained models. The pre-trained models will be evaluated comprehensively on multiple tasks with different domains. The CafeBERT model will be applied to many tasks for Vietnamese to improve performance and get many applications in the field of natural language processing in Vietnamese. In addition, resource-poor languages can monitor and work our way up to creating great pre-training models that can enhance performance and have many real-world applications.

Limitations

We have shown that the CafeBERT model achieves SOTA results on the VLUE benchmark. However, more experiments and analysis are still needed to clarify and better understand the impact of our model on tasks of the VLUE benchmark. In addition, more tests are needed for tasks other than the VLUE benchmark to clarify and understand the new model across domains and different types of tasks in Vietnamese. We leave these as motivation for future studies. In addition, we choose a large data set available instead of taking advantage of a large amount of Vietnamese data from more sources because it requires a large amount of computing power and requires hardware resources.

Ethics Statement

The authors introduced the first Vietnamese language understanding evaluation (VLUE) benchmark to evaluate the power of pre-trained language models in Vietnamese. The VLUE benchmark uses five datasets for five tasks, including UIT-ViQuAD 2.0, ViNLI, VSMEC, ViHOS, and NIIVTB POS, published previously. In addition, the authors introduce the CafeBERT pre-trained model. The new model is trained based on the XLM-Roberta model with a large Vietnamese dataset, including Wikipedia and electronic news articles.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Tin Van Huynh, Kiet Van Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Sentence extraction-based machine reading comprehension for vietnamese. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II 14*, pages 511–523. Springer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333. Springer.

Phu Gia Hoang, Canh Luu, Khanh Tran, Kiet Nguyen, and Ngan Nguyen. 2023. Vihos: Hate speech spans detection for vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for

studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

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.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. New vietnamese corpus for machine reading comprehension of health news articles.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018a. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Luan Nguyen, Kiet Nguyen, and Ngan Nguyen. 2022. SMTCE: A social media text classification evaluation benchmark and BERTology models for Vietnamese. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 282–291, Manila, Philippines. De La Salle University.

Quy Nguyen, Yusuke Miyao, Ha Le, and Ngan Nguyen. 2016. Challenges and solutions for consistent annotation of Vietnamese treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1532–1539, Portorož, Slovenia. European Language Resources Association (ELRA).

Quy T. Nguyen, Yusuke Miyao, Ha T. T. Le, and Nhung T. H. Nguyen. 2018b. Ensuring annotation consistency and accuracy for vietnamese treebank. *Language Resources and Evaluation*, 52:269–315.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Tham Thi Nguyen, Sieu Khai Huynh, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment

analysis to social listening system for business intelligence.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marian Simko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. SlovakBERT: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7156–7168, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. Wikibert models: Deep transfer learning for many languages. *NoDaLiDa 2021*, page 1.

Khanh Quoc Tran, An Trong Nguyen, Phu Gia Hoang, Canh Duc Luu, Trong-Hop Do, and Kiet Van Nguyen. 2023. Vietnamese hate and offensive detection using phobert-cnn and social media streaming data. *Neural Computing and Applications*, 35(1):573–594.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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 1071–1078, Dubrovnik, Croatia. Association for Computational Linguistics.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Tuan-Vi Tran, Xuan-Thien Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. An empirical study for vietnamese constituency parsing with pre-training. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE.

Nguyen Van Kiet, Tran Quoc Son, Nguyen Thanh Luan, Huynh Van Tin, Luu Thanh Son, and Nguyen Luu Thuy Ngan. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

Kiet Van Nguyen, Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2023. Multi-stage transfer learning with bertology-based language models for question answering system in vietnamese. *International Journal of Machine Learning and Cybernetics*, 14(5):1877–1902.

Kiet Van Nguyen, Nhat Duy Nguyen, Phong Nguyen-Thuan Do, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vireader: A wikipedia-based vietnamese reading comprehension system using transfer learning. *Journal of Intelligent & Fuzzy Systems*, 1:1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Examples of Tasks in VLUE

Table 5: Examples of each task in the VLUE benchmark

Task	Samples
UIT-ViQuAD	*Sample 1* **Context**: Đầu những năm 2000, trong Moulin Rouge! (2001), Nicole Kidman vào vai cô ca sĩ Satine của quán Moulin Rouge yêu chàng nhà văn Christian do Ewan McGregor diễn. [...] (*In the early 2000s, in the Moulin Rouge! (2001), Nicole Kidman plays Moulin Rouge singer Satine who falls in love with Christian writer Ewan McGregor.*) **Question**: Ca sĩ Satine trong phim Moulin Rouge! do ai thủ vai? (*Singer Satine in the movie Moulin Rouge! played by who?*) **Answer**: Nicole Kidman *Sample 2* **Context**: Đầu thế kỉ 20, Puerto Rico nằm dưới sự cai trị của quân đội Mỹ và thống đốc Puerto Rico đều là người được Tổng thống Mỹ chỉ định. [...] (*In the early 20th century, Puerto Rico was under the rule of the US military and the governor of Puerto Rico was both appointed by the US President.*) **Question**: Sang thế kỉ XX, cường quốc nào kiểm soát Puerto Rico? (*In the twentieth century, which country controlled Puerto Rico?*) **Answer**: Mỹ (*The US*)
ViNLI	*Sample 1* **Premise**: Rau sam trắng mọc nhiều ở ven bờ ruộng, vùng ven biển. (*White purslane grows a lot in the fields and coastal areas.*) **Hypothesis**: Chúng ta có thể dễ dàng tìm thấy rau sam trắng các vùng ven bờ ruộng hay ven biển. (*We can easily find white purslane in areas along the fields or along the coast.*) **Label**: Entailment *Sample 2* **Premise**: Ngoại trưởng Blinken tuyên bố Mỹ sẽ không để Australia một đối mặt với áp lực kinh tế từ Trung Quốc. (*Foreign Minister Blinken said the US would not leave Australia alone to face economic pressure from China.*) **Hypothesis**: Mỹ và Australia đã đồng hành cùng nhau trong công cuộc phát triển kinh tế nhiều thập niên qua. (*The US and Australia have been together in economic development for decades.*) **Label**: Neutral
VSMEC	*Sample 1* **Sentence**: lại là lào cai , tự hào quê mình quá :)) (*It's Lao Cai again, so proud of my hometown* :))) **Label**: Enjoyment *Sample 2* **Sentence**: per đúng rồi , không muốn xa cách đâu (*per is right, don't want to be far away*) **Label**: Sadness
ViHOS	*Sample 1* **Text**: Ba khùng nữa rồi (*you are crazy again*) **Label**: O B-T O O *Sample 2* **Text**: Thời trang mà dell ra gì. (*Fashion for nothing*) **Label**: O O O B-T O O
NIIVTB POS	*Sample 1* **Text**: Mọi người ồn_ào đếm tiền , ký sổ ... (*People were noisy counting money, signing books...*) **Label**: Nw Nn Aa Vv Nn PU Vv Nn PU *Sample 2* **Text**: " Chiếm rồi họ canh còn kỹ hơn bảo_vệ của công_ty", anh Vỹ kể. (*"After taking possession, they guarded more carefully than the company's security", Mr. Vy said.*) **Label**: PU Vv R Pp Vv R Aa Vcp Nn Cs Nn PU PU Nn Nr Vv PU