Error Investigation of Pre-trained BERTology Models on Vietnamese Natural Language Inference

Tin Van Huynh[1,2(✉)], Huy Quoc To[1,2], Kiet Van Nguyen[1,2], and Ngan Luu-Thuy Nguyen[1,2]

[1] Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh, Vietnam
{tinhv,huytq,kietnv,ngannlt}@uit.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Natural Language Inference tasks have emerged in recent years and attracted significant attention from the natural language processing research community. There has been much success in this task with many quality datasets in English and Chinese for research and demonstrating the impressive performance of machine learning models. Pre-trained models play a crucial role, which is reflected in their superior performance compared to other models. However, they are still far from perfect and have many obstacles to the characteristics of the data. Especially in Vietnamese, we have just seen the emergence of the ViNLI benchmark dataset to serve the research community. In this paper, we experiment and analyze how the characteristics in the ViNLI benchmark dataset affect the performance of the pre-trained BETology-based models. In addition, the data parameters of ViNLI are also measured and analyzed on the accuracy of these models to see if it has any impact on the accuracy of the model.

Keywords: Natural language inference · Error analysis

1 Introduction

The original NLI task, known as Recognizing textual entailment [8,9], required the machine learning model to capture the semantics of a given pair of premise and hypothesis sentences. This semantic relationship can fall into cases like Entailment, Contradiction, or Neutral. In recent years, the Natural Language Inference task has achieved significant success, which plays a crucial role because it affects many NLP tasks such as machine reading comprehension [18] and question answering [4]. The remarkable point in this task is that the presence of many high-quality large datasets in many different languages ranging from rich-resource languages such as English [2,20,26] and Chinese [15] to poor-resource languages such as Korean [14], Indonesian [17], and Persian [1]. As a low-resource language, Vietnamese still has many limitations for outstanding research in this NLI task. However, recently the research community has witnessed the launch of the ViNLI dataset, which was developed by Huynh et al. [16] for Vietnamese. This dataset has yielded some positive research results, so it is hoped to promote more and better research outcomes in the future.

E. Szczerbicki et al. (Eds.): ACIIDS 2022, CCIS 1716, pp. 176–188, 2022.
https://doi.org/10.1007/978-981-19-8234-7_14

It can be seen that there is an interplay between datasets and machine learning models. In other words, datasets play an essential role in the evaluation of machine models, and machine learning models are increasingly thriving to improve accuracy on NLI tasks dramatically. In particular, the appearance of transformer architecture [24] is a leap forward for developing various tasks in NLP, including NLI task. After that, the BERTology model [10] is becoming a trend thanks to its transformer-based architecture. However, we still do not fully understand why BERT has such good performance, which is also a problem for that many researchers are trying to find an explanation.

In this paper, we try to investigate the behavior of the pre-trained BERT language model and variant models of BERT through the lens of the Vietmanses NLI task. Vietnamese is an interesting language, but not much research has been done. From the current research results from the ViNLI dataset [16], we focused on setting up experiments in this paper. We deeply analyzed the features contained in ViNLI to see what affects the pre-trained model performance. This study can help us better understand pre-trained models as well as the ViNLI dataset. We hope these analyses point to potential future studies to improve the Vietnamese NLI task outcomes further.

2 Related Work

In recent years, many NLI datasets have been built for studying the effectiveness of machine learning models such as deep learning and transfer learning. Many large benchmark datasets have been introduced related to human natural language inference. Specifically, the dataset named SNLI [2] introduced in 2015 is a large manually labeled dataset from Stanford University. Then, a series of other datasets appeared, such as STS-B [3], QQP [5], introduced in 2017 and 2018 for English. In 2018, a large dataset for this language was also published for research as MultiNLI [26] with 433K pairs. In addition, datasets for various languages have emerged in the NLP Research communities, including FarsTail [1] for Persian, KorNLI & KorSTS [14] for Korean, IndoNLI [17] for Indonesian, and OCNLI [15] for Chinese. Regarding the multilingual dataset, the XNLI dataset [7] was released in 2018 with more than 112K pairs for 15 languages. In Vietnamese, we have a ViNLI dataset introduced by Huynh et al. [16] to promote NLI research in Vietnamese.

Natural language inference research is growing rapidly due to the explosion of high-quality large datasets and deep learning models. Besides machine learning models based on neural networks such as RNN [11], Bi-LSTM [12] have achieved good performance on this task, the transformer-based core model plays a vital role. BERT was published by Devlin et al. [10]. Its architecture includes a variable number of Transformer encoder layers and self-attention heads. With this architecture, BERT achieves many state-of-the-art results for several Natural Language Understanding tasks on different datasets such as GLUE benchmark [25], SQuAD [22], and SWAG [28]. With the NLI task, pre-trained transformer models on many languages such as multilingual BERT [10], XLM-R [6], SBERT [23] give surprising results on the datasets MultiNLI [26], XNLI [7], QQP [5], STS-B [3]. PhoBERT [19] is a monolingual pre-trained model developed only for Vietnamese that is also giving positive results on many NLP tasks such as text classification, natural language inference, or named entity recognition.

3 Dataset

The ViNLI benchmark [16] is used for evaluating the accuracy of pre-trained models. The statistics on the dataset are shown in Table 1. ViNLI is an open-domain dataset built on Vietnamese news text. This dataset is quite large for Vietnamese at the moment, with 30,376 pairs of premise-hypothesis sentences manually annotated by humans. The special thing about ViNLI compared to other datasets is that it has an additional label Other instead of three labels Entailment, Contradiction, and Neutral like other datasets. The authors added the Other label to distinguish it from the Neural label.

Table 1. The number of premises-hypothesis pairs in the ViNLI dataset.

Label	Quantity			
	Train	Dev	Test	Total
Entailment	6,094	739	750	7,583
Contradiction	6,094	764	737	7,595
Neutral	6,094	752	777	7,623
Other	6,094	754	727	7,575
Total	**24,376**	**3,009**	**2,991**	**30,376**

4 Experiments and Results

This section presents experiments with multilingual pre-trained models on the ViNLI dataset. Following the prior work [2,16], we use the accuracy measures and F1-score to evaluate the performance of those models.

4.1 Data Preparation

The ViNLI benchmark dataset is used for experiments on pre-trained models. However, according to the experimental results of Huynh et al. [16], the accuracy of the best model giving accurate results on the Other label is very high, above 98%, so we focus on the analysis of the dataset with three labels Contradiction, Entailment, and Neutral. Therefore, before installing the experiment, we remove the pairs of sentences labeled Other from the train, dev, and test set.

4.2 Experiment Settings

Besides experiments with pre-trained models, including multilingual BERT [10], PhoBERT [19], XLM-R [6] established on ViNLI by Huynh et al. [16], we also carry out the experiment on a model Another pre-trained model is SBERT [23]. The SBERT model is pre-trained in many different languages, including Vietnamese. We use these pre-trained models provided to HugggingFace's library in our experiments. The parameters in the SBERT model are we set up as follows: learning_rate = 1e−05, batch_size = 16, max_length = 256, in addition, we set epoch = 10.

4.3 Experimental Results

The experimental results are shown in Table 2. Compared with the experimental results of Huynh et al. [16], it can be seen that the performance of the SBERT model is the lowest with the accuracy on the dev and test sets of 59.29% and 58.17%, respectively. Besides, the experimental results on SBERT have a rather large gap compared with other pre-trained models, especially when compared with the XLM-R_large model. This difference in accuracy is more than 23% on both the dev set and test set.

Table 2. Machine performances on the development and test sets of ViNLI dataset. Results of mBERT, PhoBERT, and XLM-R are from Huynh et al. [16]

Model	Dev		Test	
	Acc	F1	Acc	F1
SBERT$_{Base}$	57.83	57.85	57.33	57.32
SBERT$_{Large}$	59.29	59.03	58.17	57.69
mBERT	67.41	67.46	64.84	64.83
PhoBERT$_{Base}$	75.07	75.08	72.87	72.79
PhoBERT$_{Large}$	**77.33**	**77.34**	**75.93**	**75.87**
XLM-R$_{Base}$	72.02	71.99	71.59	71.51
XLM-R$_{Large}$	**83.02**	**82.98**	**81.36**	**81.31**

5 Result Analysis

In this section, we carry out an analysis of the results of these pre-trained models, which aims to explore how the characteristics of the ViNLI dataset affect the performance of these pre-trained models. The issues in ViNLI that we are interested in analyzing include the influence of the annotation rule, word overlap, sentence length on performance, ability to capture annotation artifacts of pre-trained models, and error analysis by confusion matrixes.

5.1 Effects of Annotation Rules

According to Huynh et al. [16], to build the ViNLI dataset, annotators have to follow an annotation guideline. In the guidelines, they present suggested rules for annotators to writing a hypothesis corresponding to a premise sentence. To analyze how the characteristics of the ViNLI construction method affect the results of the pre-trained models, we investigate how the rules of creating hypothesis sentences for entailment and contradiction labels affect the performance of models. The rules list for creating the hypothesis sentences of the label entailment and contradiction is shown in Table 3 and Table 4. We selected 200 premise-hypothesis pairs of the entailment label and 200 premise-hypothesis pairs of the contradiction label in the test set for analysis. From these 400 pairs of sentences, we annotate the creating hypothesis sentence rules for these pairs of

sentences following guidelines of Huynh et al. [16]. The percentages of each rule generating the hypothesis of the label entailment and contradiction are shown in Table 3 and Table 4, respectively.

In terms of entailment rules, we found that annotators tended to use the "replace words with synonyms" rule the most, with 56%. Besides, rules like "Add or remove modifiers that do not radically alter the meaning of the sentence" and "Change active sentences into passive sentences and vice versa" also account for a significant percentage of the annotators' writing style, with 54% and 35%, respectively. In contrast, rules like "Turn adjectives into relative clauses", "Create conditional sentences", or "Turn the object into relative clauses" are the least used by annotators to create the entailment hypothesis, with only from 1% to less than 4%.

We observe that the accuracy results of the pre-trained models on the entailment rules in Table 3 are interesting, with many similarities and differences between the models. All four models, SBERT, mBERT, PhoBERT, and XLM-R have the worst performance on pairs of sentences generated from the rule "Turn adjectives into relative clauses" even the mBERT model does not correctly predict any pairs, while the other three models correctly predicted half of those pairs of sentences. In addition, the rule "Create conditional sentences" is also a rule that makes it difficult for mBERT, PhoBERT, and XLM-R models with lower accuracy compared to the accuracy of other rules. SBERT model has the highest accuracy on two rules, "Replace words with synonyms" and "Create conditional sentences" with over 66%. Furthermore, the PhoBERT model has the best performance on the pairs of entailment sentences generated from the rule "Add or remove modifiers that do not radically alter the meaning of the sentence" with 86.11%. Both the mBERT and XLM- R models have the highest accuracy on the rule "Turn the object into relative clauses" with 85.71% and 100%, respectively.

Table 3. Statistics of rules generate entailment sentences and the accuracy of pre-trained models on these rules.

No.	Rule	Occurrence percentage (%)	Accuracy (%)			
			SBERT	mBERT	phoBERT	XLM-R
1	Change active sentences into passive sentences and vice versa	35.0	60.00	65.71	81.43	91.43
2	Replace words with synonyms	**56.0**	**66.96**	62.50	82.14	91.07
3	Add or remove modifiers that do not radically alter the meaning of the sentence	54.0	62.96	63.89	**86.11**	92.59
4	Replace Named Entities with a word that stands for the class	13.5	59.26	62.96	85.18	92.59
5	Turn nouns into relative clauses	4.0	62.50	37.50	75.00	75.00
6	Turn the object into relative clauses	3.5	57.14	**85.71**	71.43	**100.00**
7	Turn adjectives into relative clauses	**1.0**	**50.00**	**0.00**	**50.00**	**50.00**
8	Replace quantifiers with others that have a similar meaning	11.5	56.52	60.87	78.26	91.30
9	Create a presupposition sentence	12.5	56.00	64.00	80.00	80.00
10	Create conditional sentences	**1.5**	**66.67**	33.33	66.67	66.67
11	Other	11	**50.00**	63.63	68.18	81.81

Regarding contradiction rules, Annotators frequently use the "Replace words with antonyms" rule to generate most hypothesis sentences with over 36%, around five times as high as the "Opposite of time" rule, which has the lowest percentage. In addition, the percentage of contradiction hypothesis generated from "Opposite of quantity" and "Opposite of time" rules is quite low. Table 4 shows that four pre-trained models have the best predictive ability on pairs of "Use negative words" rule with high accuracy, especially mBERT and XLM-R models achieve nearly 90%. In particular, the XLM-R model does not have difficulty with pairs of sentences belonging to "Other" rules with absolute accuracy up to 100%, while the number of these pairs of sentences in the dataset is the lowest. The analysis results also show that the SBERT model has the worst performance on the hypothesis sentences generated from the "Replace words with antonyms" rule with only 32.87%. In contrast, the predictive ability of the PhoBERT and XLM-R model on this rule is quite high relative to 82.19% and 87.67%. Besides, The mBERT model has the lowest accuracy on sentence pairs from the rule "Wrong reasoning about an event". PhoBERT's accuracy is the lowest on the "Opposite of time" rule with around 50%.

Table 4. Statistics of rules generate contradiction sentences and the accuracy of pre-trained models on these rules.

No.	Rule	Occurrencepercentage (%)	Accuracy (%)			
			SBERT	mBERT	phoBERT	XLM-R
1	Use negative words (no, not, never, nothing, hardly, etc.)	19.5	**76.92**	**89.74**	**84.61**	**89.74**
2	Replace words with antonyms	**36.5**	**32.87**	61.64	82.19	87.67
3	Opposite of quantity	9.0	72.22	66.67	83.33	**72.22**
4	Opposite of time	7.0	64.28	64.29	**50.00**	78.57
5	Create a sentence that has the opposite meaning of a presupposition	23.5	51.06	57.45	70.21	**72.34**
6	Wrong reasoning about an object (House, car, river, sea, person, etc.)	19.5	46.15	66.67	66.67	74.36
7	Wrong reasoning about an event	20.5	46.34	**51.22**	65.84	80.49
8	Other	2.5	40.00	60.00	80.00	**100.00**

We also analyze how annotators combine multiple rules to write hypothesis statements that affect the performance of pre-trained models. The ratio of the number of rules have in a hypothesis is shown in Table 5, along with the performance of the pre-trained models. In general, the number of rules used to generate the entailment hypothesis sentences is equally distributed over 1, 2, and more than 2 rules. In addition, most contradiction hypothesis sentences are written using a rule with 63% and a lower percentage of 37% for cases generated from more than 1 rule. We observe on the entailment label that while mBERT has the best accuracy on the entailment hypothesis sentences with only 1 rule, accuracy decreases as the number of rules increases. In contrast, the performance of the XLM-R model increases as the number of rules used to generate the entailment hypothesis sentences increases. Besides, both SBERT and PhoBERT models have the best predictive ability on entailment hypothesis sentences with 2 rules and

maintain stability with 1 or more than 2 rules. For the number of rules in the contradiction hypothesis, all four pre-trained models have better predictability when the hypothesis is generated from multiple rules.

Table 5. The effect of the number of rules in the entailment and contradiction hypothesis sentence on the performance of the pre-trained models.

Label	Number of rule	Occurrence percentage (%)	Accuracy (%)			
			SBERT	mBERT	phoBERT	XLM-R
Entailment	1 rule	30.5	60.66	**72.13**	80.33	81.96
	2 rules	39.5	**64.55**	63.29	**83.54**	88.61
	More than 2 rules	30.0	60.00	58.33	80.00	**93.33**
Contradiction	1 rule	63.0	41.27	61.11	69.84	77.78
	More than 1 rules	37.0	**58.11**	**78.38**	**78.38**	**83.78**

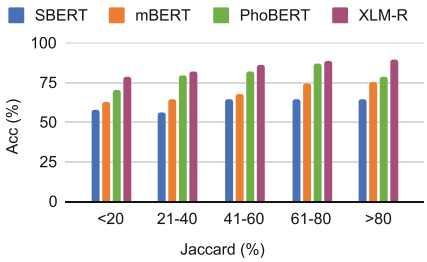
5.2 Effects of Word Overlap

To analyze whether word overlap between premise and hypothesis sentences in ViNLI affects the performance of pre-trained models? We calculate the word overlap of premise-hypothesis pairs on the test set according to three different metrics, including Jaccard, The Longest Common Subsequence (LCS), and new token rate similar [17]. And then, we analyze the accuracy of the models according to these measures.

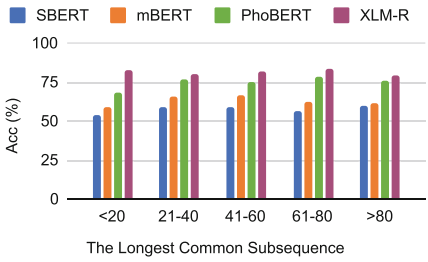
First, we use Jaccard to measure the degree of unordered word overlap by token level; The resulting accuracy of models by Jaccard is shown in Fig. 1a. It can be seen that the XLM-R model has the best performance on all Jaccard ranges. The accuracy of the SBERT model is quite low when Jaccard is less than 40%, and then slightly increases with Jaccard in a range of from 42% to more than 80%. All three models, mBERT, PhoBERT, and XLM-R have the worst performance when the Jaccard between the premise and the hypothesis is less than 20%, and performance increases dramatically as the Jaccard increases. However, the accuracy of the PhoBERT model decreases significantly when the Jaccard between premise and hypothesis sentences is more than 80%.

Second, we use LCS to measure the degree of word overlap in order between the premise and the hypothesis sentences by character. The accuracy of the models according to the LCS is indicated in Fig. 1b. We found that the XLM-R model has the highest performance and is relatively stable on most levels of LCS compared to the other models. While the PhoBERT model has low performance on premise-hypothesis pairs with LCS less than 20 characters, the mBERT model has difficulty when sentence pairs have LCS less than 20 characters and higher than 60 characters.

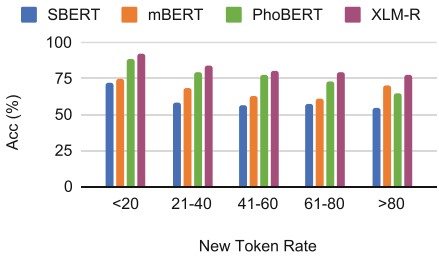
Third, we also analyze the results of the models according to the ratio of new words in the hypothesis sentence compared to the premise sentence. The analysis results are shown in Fig. 1c. Most of the performance of pre-trained models decreases remarkably as the new word rate increase from 0 to more than 80%.



(a) Accuracy on Jaccard.



(b) Accuracy on LCS.



(c) Accuracy on new token rate.

Fig. 1. The effect of word overlap on the accuracy of pre-trained models.

From these analysis results, it can be seen that the degree of word overlap between the premise and the hypothesis sentences significantly influences the accuracy of the pre-trained models.

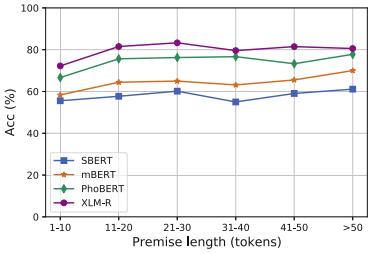
5.3 Effect of Sentence Length

The issue we are also interested in analyzing in this section is the effect of the length of inference sentences pair on the performance of pre-trained models. Models' accuracy on the test set concerning the length of the premise sentence, the length of the hypothesis sentence, and the total length of the premise sentence and the hypothesis sentence by token are shown in Figs. 2a, 2b, and 2c, respectively. We found that the accuracy of most models increases significantly as the length of the premise sentences rises from 1–10 tokens to 21–30 tokens. While the accuracy of the PhoBERT and mBERT models continues to increase slightly as the premise sentence length rises to more than 50 tokens, the XLM-R and SBERT models decrease slightly.

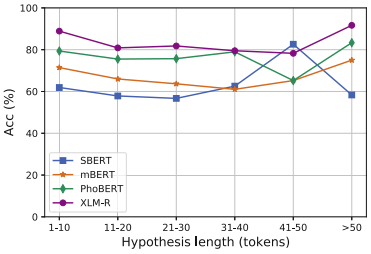
Regarding the hypothesis sentence length, the mBERT and XLM-R model's accuracy decreases significantly when the hypothesis sentence length increases from 1 to 40 tokens, followed by a gradual escalation when the hypothesis sentence length is more than 40 tokens. Looking at the 2b figure, we find that the performance of the PhoBERT and SBERT model when the same when the hypothesis sentence length is in the range of 1 to 40 tokens. While SBERT's performance continued to surge above 80% when the length of the hypothesis sentence increased from 41–50 Tokens before its perfor-

mance dropped below 60% when the hypothesis length sentence length was more than 50 tokens if we take a look at PhoBERT models, we will see an opposite trend.

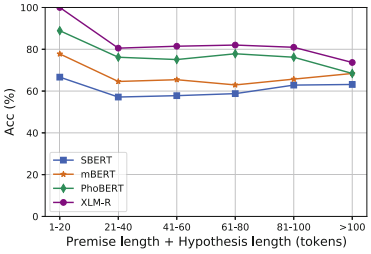
We find that the performance of the SBERT, mBERT, PhoBERT, and XLM-R model is relatively high when the total length of the premise and hypothesis is between 1–20 tokens; even the XLM-R model is almost entirely correct. However, the performance of these models goes down significantly as this total length increases from 20 tokens to more than 100 tokens.



(a) Accuracy on premise sentence length.



(b) Accuracy on hypothesis sentence length.



(c) Accuracy on the total length of the premise and hypothesis sentences.

Fig. 2. The effect of length of premise and hypothesis sentences on pre-trained models.

5.4 Hypothesis only Model Analysis

Inspired by the research of [21], we investigate whether the annotation artifacts leave any clues on the hypothesis sentence that help language inference models correctly predict the label. The models' performance is trained with only hypotheses illustrated in Table 6. We observe that the XLM-R and PhoBERT models have pretty impressive results when the accuracy on the Test set with 56.63% and 57.68%, respectively. Besides, we calculate Pointwise Mutual Information (PMI) [13] to observe which words in the hypothesis sentences can distinguish labels from each other. PMI results for the top 5 words of each label are shown in Table 7. With the entailment label, we found it quite interesting that the word *"không"* is actually a word that represents this class.

Table 6. Hypothesis-only baselines for ViNLI.

Model	Dev		Test	
	Acc	F1	Acc	F1
SBERT	50.51	50.48	49.91	49.88
mBERT	52.82	52.72	53.48	53.35
PhoBERT	56.14	56.12	57.68	57.59
XLM-R	57.87	57.67	56.63	56.43

This is entirely different from the OCNLI [15] and IndoNLI [17] datasets, where negative lexical dominate in hypothesis sentences of contradiction label. In addition, the word *"có"* and *"một"* can be a sign to discriminate the neutral class from other classes. However, the PMI results also show that some words can represent multiple classes, such as *"và"* and *"trong"*. There's not too influential in terms of the lexical difference between classes. Therefore, pre-trained models are made difficult by the ViNLI dataset if only trying to rely on hypothesis sentences to predict.

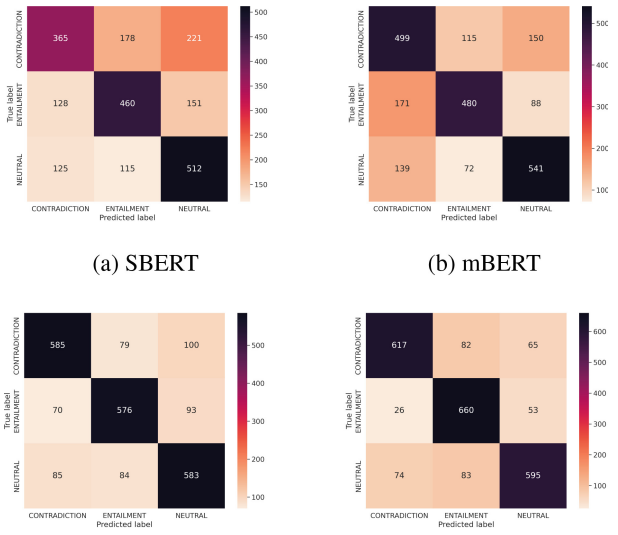
Table 7. Top 5 (word, label) pairs PMI for different labels of ViNLI.

Word		**Label**	**PMI**	**Percentage**
và	*and*	E	0.18	17.17
các	*some*	E	0.22	15.13
của	*of/object's*	E	0.28	27.51
trong	*in/inside*	E	0.35	18.94
không	*no/not*	E	0.35	9.81
của	*of/object's*	C	0.19	24.26
là	*to be*	C	0.21	16.19
trong	*in/inside*	C	0.22	15.73
và	*and*	C	0.23	18.45
các	*some/several*	C	0.30	17.17
một	*one/a/an*	N	0.27	10.49
là	*to be*	N	0.31	18.79
và	*and*	N	0.32	21.13
có	*has/have*	N	0.33	18.83
trong	*in/inside*	N	0.33	18.45

5.5 Error Analysis by Confusion Matrixes

Figure 3 illustrates the confusion matrix of the four pre-trained models the development set, including SBERT, mBERT, PhoBERT, and XLM-R. While the SBERT, mBERT, and PhoBERT models erroneously predict a significant number of sentence pairs with the CONTRADICTION label to the NEUTRAL label, many contradictory sentence

pairs are mistakenly predicted by the XLM-R model as the label ENTAILMENT. In addition, the rate of EMTAILMENT sentence pairs being mispredicted to the CONTRADICTION label and the NEUTRAL label was quite similar for each model except for the mBERT model, which had more false predictions to the CONTRADICTION label than to the NEUTRAL label. With sentence pairs of the NEUTRAL label, the XLM-R model has the best prediction ability on this label. Meanwhile, the mBERT model gives a significantly incorrect prediction from the NEUTRAL label to the CONTRADICTION label.



(a) SBERT (b) mBERT

(c) PhoBERT (d) XLM-R

Fig. 3. Confusion matrix of pre-trained language models on the development set.

6 Conclusion and Future Work

To analyze the performance of pre-trained models on the Vietnamese NLI task, we experimented with the SBERT model on the ViNLI dataset and in-depth analysis of other pre-trained models experimented by Huynh et al. [16]. There are many interesting findings relating between data characteristics and the accuracy of models. In particular, most models have relatively low accuracy on the sentences entailment hypothesis generated from the rules “Turn adjectives into relative clauses” and “Create conditional sentences”. The contradiction hypothesis generated from the “Use negative words” rule is straightforward for the models to predict correctly. In addition, when multiple rules are combined to create a contradiction hypothesis, the prediction models are more accurate. Word overlap or premise and hypothesis length also significantly affect the model’s

performance. Pre-trained models are able to make predictions thanks to the clues of the annotation artifacts, although the accuracy is not too high.

In the future, we will also learn techniques to improve the accuracy of the models, such as data enhancement techniques. Besides, we will explore other transformer models like mT5 [27] which is a pre-trained text-to-text transformer in many languages.

Acknowledgement. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2022-26-01. Tin Van Huynh was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2021.ThS.49.

References

1. Amirkhani, H., et al.: FarsTail: a Persian natural language inference dataset. arXiv preprint arXiv:2009.08820 (2020)
2. Bowman, S., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642 (2015)
3. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.:. SemEval-2017 task 1: semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055 (2017)
4. Chen, J., Choi, E., Durrett, G.: Can NLI models verify QA systems' predictions? In: Findings of the Association for Computational Linguistics, EMNLP 2021, pp. 3841–3854 (2021)
5. Chen, Z., Zhang, H., Zhang, X., Zhao, L.: Quora question pairs (2018). https://www.kaggle.com/c/quora-question-pairs
6. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale (2019). arXiv preprint arXiv:1911.02116
7. Conneau, A., et al.: XNLI: evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2475–2485 (2018)
8. Cooper, R., et al.: Using the framework (1996)
9. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177–190. Springer, Heidelberg (2006). https://doi.org/10.1007/11736790_9
10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
12. Ghaeini, R.: Dependent reading bidirectional LSTM for natural language inference. arXiv preprint arXiv:1802.05577 (2018)
13. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. arXiv preprint arXiv:1803.02324 (2018)
14. Ham, J., Choe, Y.J., Park, K., Choi, I., Soh, H.: KorNLI and korSTS: new benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289 (2020)
15. Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., Moss, L.S.: OCNLI: original Chinese natural language inference. In: Findings of the Association for Computational Linguistics, EMNLP 2020, pp. 3512–3526 (2020)

16. Van Huynh, T., Van Nguyen, K., Nguyen, N.L.-T.: ViNLI: a Vietnamese corpus for studies on open-domain natural language inference. In: Proceedings of the 29th International Conference on Computational Linguistics (Accepted) (2022)
17. Mahendra, R., Aji, A.F., Louvan, S., Rahman, F., Vania, C.: IndoNLI: a natural language inference dataset for Indonesian. arXiv preprint arXiv:2110.14566 (2021)
18. Mishra, A., Patel, D., Vijayakumar, A., Li, X., Kapanipathi, P., Talamadupula, K.: Reading comprehension as natural language inference: a semantic analysis. In: Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, pp. 12–19 (2020)
19. Nguyen, D.Q., Nguyen, A.T.: PhoBERT: pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020, 1037–1042 (2020)
20. Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., Kiela, D.: Adversarial NLI: a new benchmark for natural language understanding. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4885–4901 (2020)
21. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B.: Hypothesis only baselines in natural language inference. arXiv preprint arXiv:1805.01042 (2018)
22. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392 (2016)
23. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
25. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: 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, pp. 353–355 (2018)
26. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)
27. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. In: NAACL-HLT (2021)
28. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: A large-scale adversarial dataset for grounded commonsense inference. In: EMNLP, Swag (2018)