ViLegalNLI: Vietnamese NLI dataset for Legal Domain

Nguyễn Phi Long[1], Hồ Nguyễn Thiên Vũ[1], Dương Thị Hồng Nhung[1]
[1]University of Information Technology, VNU-HCM
22520818@gm.uit.edu.vn, 22521689@gm.uit.edu.vn, 22521056@gm.uit.edu.vn

Abstract

Over the past decade, the field of computational linguistics has seen significant growth in the development of datasets and models for Natural Language Inference (NLI) for resource-rich languages like English and Chinese. However, for Vietnamese, a low-resource language, creating a large-scale, high-quality dataset for NLI is essential. In this paper, we introduce ViLegalNLI (Vietnamese Legal Natural Language Inference), a high-quality dataset for legal texts designed to evaluate Vietnamese NLI models. This dataset was created and validated through a strict quality control process. ViLegalNLI consists of over 9,600 premise-hypothesis sentence pairs, with hypotheses generated by LLM models, extracted from more than 3,000 legal documents across 23 different domains. This paper also outlines specific guidelines for corpus creation, taking into account the unique linguistic characteristics of Vietnamese in expressing entailment and contradiction. To assess the difficulty of the dataset, we conducted experiments with popular NLP methods and state-of-the-art pre-trained models. The best system performance exceeded expectations, achieving an accuracy of over 92%. The ViLegalNLI corpus not only serves as a significant challenge for research in Vietnamese computational linguistics but also contributes to advancing NLI model development for low-resource languages.

1 Introduction

Natural Language Processing (NLP) plays a crucial role in enabling computers to understand and analyze human language, thereby driving practical applications across various fields. NLP supports virtual assistant systems, machine translation, text summarization, document classification, and enhances communication experiences through chatbots. It also facilitates sentiment analysis, public opinion mining, and optimization of information retrieval systems. In education and research, NLP contributes to linguistic analysis, grammar checking, and automatic content generation, promoting academic research and improving training quality (Jurafsky, D., & Martin, J. H., 2023).

Core NLP tasks like Natural Language Inference (NLI) have seen significant advancements, fueled by multilingual datasets such as SNLI [2], MultiNLI [3], and XNLI [4], alongside state-of-the-art deep learning models like BERT [5] and XLM-R [6]. NLI, a fundamental task in natural language understanding, requires predicting the semantic relationship between two sentences (entailment, contradiction, or neutral). Recent efforts have expanded NLI research beyond English to other languages, with datasets such as OCNLI (Chinese) [6], KorNLI (Korean) [7], and IndoNLI (Indonesian) [8], advancing multilingual and low-resource NLP studies.

Although Vietnamese is spoken by over 90 million people, it is still considered a low-resource language in NLP research, particularly for complex tasks like NLI. The lack of diverse and high-quality datasets poses significant challenges for developing effective NLI models for Vietnamese. However, recent initiatives have focused on building Vietnamese datasets, such as ViNLI, to facilitate the evaluation and development of more robust NLI models for this language [9][10]. Despite these efforts, legal text datasets in Vietnamese remain underexplored and have not yet been fully developed to address the NLI task. Furthermore, previous studies on NLI have been limited to the sentence-level and have not extended to the document-level.

To address this gap, we introduce ViLegalNLI, a

dataset designed for the NLI task on Vietnamese legal texts. This dataset comprises 9,667 hypothesis-premise pairs, derived from 3,235 legal documents spanning 23 different domains, collected from legal websites. ViLegalNLI aims to advance the evaluation of NLI models and enhance NLP research for Vietnamese legal texts.

Our main contributions are summarized as follows:

- We introduce ViLegalNLI, a dataset of Vietnamese legal texts containing 9,667 samples, designed to evaluate the performance of models on Natural Language Inference (NLI) tasks for legal documents.

- We conduct experiments on Transformer-based models pre-trained on both multilingual and Vietnamese-specific datasets.

- We perform an in-depth analysis to explore the impact of exploratory data analysis on experimental results, providing insights into the strengths and limitations of the dataset.

2 Related Work

The Natural Language Inference (NLI) task has garnered significant attention from the international research community. One of the most prominent contributions is the Stanford Natural Language Inference (SNLI) Corpus by Bowman et al. (2015) [2]. SNLI provides a dataset of 570,152 sentence pairs with three main labels: entailment, contradiction, and neutral, serving as a benchmark for training deep learning models. However, SNLI primarily focuses on simple conversational language, leaving complex natural language scenarios underrepresented.

In the scientific domain, SciNLI by Sadat and Caragea (2022) [11] expanded NLI research into specialized contexts. This dataset comprises 107,412 sentence pairs extracted from scientific papers, enabling the evaluation of models' ability to process scientific language. Nevertheless, challenges persist in addressing the complexity and achieving accuracy on such data. Another notable contribution is NeuralLog by Chen et al. (2021) [12], which integrates deep learning with logical reasoning to improve NLI results on datasets like SICK [13] and MED [14]. While effective, this approach requires substantial computational resources and is complex to implement, posing challenges for practical deployment.

In Vietnam, NLI has also received increasing attention with notable contributions. ViNLI by Quyen, L. H., et al. (2022) [10] introduced the first Vietnamese dataset, containing over 30,000 sentence pairs from online articles spanning 13 topics. ViNLI provides a quality resource for training and evaluating NLI models in the Vietnamese context. However, its major limitation lies in the lack of diversity in complex scenarios and the scarcity of linguistic resources compared to other languages. Recently, Huynh et al. (2024) [15] developed ViANLI, an adversarial NLI dataset with 10,000 sentence pairs designed to challenge existing models. Experimental results show that even advanced models achieved only 48.4% accuracy, highlighting the difficulty of ViANLI. However, constructing adversarial data is resource-intensive and demands high precision in annotation. Furthermore, Huyen et al. (2024) [16] introduced ViHealthNLI, an NLI dataset for the healthcare domain, paving the way for practical applications in medical information processing. However, applying NLI in healthcare requires extremely high accuracy and rigorous validation to ensure safety and effectiveness.

Despite these advancements, several limitations remain. Current datasets do not fully encompass complex contexts, particularly in specialized domains such as science and healthcare. Building large-scale datasets is resource-intensive, often leading to issues with data quality and label inconsistency. Additionally, existing NLI models struggle with complex logical reasoning and integrating information from multiple sources. Addressing these challenges is critical for advancing NLI research and applications.

3 Create Corpus

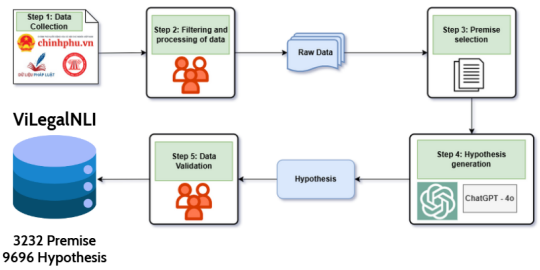

Figure 1: Dataset Construction Process

The dataset construction process is illustrated in Figure 1, comprising the following steps: Data Collection, Premise Selection, Hypothesis Generation, and Data Validation.

3.1 Data Collection

We utilized the BeautifulSoup and Selenium tools to collect data from official and reliable legal websites in Vietnam, such as Thư viện Pháp luật, Dữ liệu Pháp luật, and Văn bản Chính phủ, obtaining over 16,000 legal documents. Subsequently, we manually filtered the documents to construct the dataset based on the following criteria:

- Complete and detailed legal content.

- Accompanied by supplementary documents.

- Belonging to the types of documents: Decisions, Circulars, Decrees, Resolutions, Plans, and Directives. - Not overly lengthy (under 10,000 words).

3.2 Premise Selection

Since the processing is at the document level, our Premises consist of individual paragraphs. After filtering the data, for documents exceeding 500 tokens in length, we extract 500 tokens from random start or end positions. To ensure the paragraph is meaningful and captures the maximum content of the document, we remove any truncated words between sentences. This process also ensures that the input remains within the token limits of large language models (LLMs). As a result, from 3,232 legal documents, we extracted a corresponding 3,232 Premises.

3.3 Hypothesis Generation

With 3232 Premises obtained from the previous step, we input them into LLM models to generate 3 hypotheses for each premise, corresponding to 3 labels: 0 (Support), 1 (Neutral), 2 (Refute), defined as follows:

- 0 (Support): A hypothesis that can be inferred from the content of the Premise.

- 1 (Neutral): The content of the Premise does not provide sufficient basis to verify the validity of this hypothesis.

- 2 (Refute): A hypothesis that contradicts the content of the Premise.

The result is a dataset containing 9696 samples.

3.4 Data Validation

After generating hypotheses, we conducted manual review and editing to ensure the quality of the final dataset through the following steps:

- Correcting spelling and formatting errors.

- Removing unnecessary content (national titles, headers, numbers, etc.).

- Verifying the accuracy and appropriateness of the hypotheses and making corrections if necessary.

- Handling missing values due to errors in the data construction process.

Finally, we produced the complete ViLegalNLI dataset with 9667 samples derived from 3232 legal documents, ready for exploratory analysis and model training.

3.5 Aspect Classification

To provide deeper and more intuitive insights into the content of legal documents in ViLegalNLI, we added an 'Aspect' attribute with classification values representing the main fields addressed in these documents. Next, we created an $'aspect_keywords'$ dictionary, where keys are fields and values are keywords corresponding to each field. The fields were compiled based on classification categories from legal websites, covering 26 aspects: "bộ máy hành chính"(administrative apparatus), "tài chính nhà nước"(public finance), "văn hóa - xã hội"(culture - society), "tài nguyên - môi trường"(resources - environment), "bất động sản"(real estate), "thương mại"(commerce), "xây dựng - đô thị"(construction - urban), "thể thao - y tế"(sports - health), "giáo dục"(education), "thuế - phí - lệ phí"(tax - fees - charges), "giao thông - vận tải"(transportation - logistics), "lao động - tiền lương"(labor - wages), "đầu tư"(investment), "công nghệ thông tin"(information technology), "doanh nghiệp"(enterprise), "xuất nhập khẩu"(import - export), "tiền tệ - ngân hàng"(currency - banking), "dân sự"(civil), "dịch vụ pháp lý"(legal services), "bảo hiểm"(insurance), "thủ tục tố tụng"(procedural law), "vi phạm hành chính"(administrative violations), "kế toán - kiểm toán"(accounting - auditing), "trách nhiệm hình sự"(criminal liability), "sở hữu trí tuệ"(intellectual property), "chứng khoán"(securities), "khác"(others). For each field, excluding "khác"(others), we created a corresponding list of keywords. These keywords were selected from the content of classified legal documents on legal websites to ensure their prevalence and recognition. For example: ["cơ quan hành chính"(administrative agency), "thủ tục hành chính"(administrative procedure), "quyết định hành chính"(administrative decision), "tổ chức bộ máy"(organizational apparatus), "cán bộ công chức"(civil servant), "chế độ công vụ"(public service regime), "giấy phép hành chính"(administrative permit), "điều hành"(administration), "chỉ đạo"(directive), "giám sát hành chính"(administrative supervision), "hành vi hành chính"(administrative act), "khiếu nại hành

chính"(administrative complaint), "tố cáo hành chính"(administrative accusation)].

To classify the fields for legal documents based on the aforementioned dictionary, we counted the occurrences of keywords from each field in the document and then selected the field with the highest keyword occurrence. If none of the keywords from the dictionary appeared in the document, we assigned it to the 'khác' field. The visual results showing the distribution of fields in the dataset are presented in Figure 6.

3.6 Corpus Analysis

We conducted an initial exploratory analysis of ViLegalNLI and obtained some notable results. Firstly, we divided the dataset into Train, Dev, and Test sets with a ratio of 8:1:1. The detailed statistical results of the data are presented in Table 1.

The table reveals several notable points about the ViLegalNLI dataset. First, the "bộ máy hành chính"(administrative apparatus) aspect dominates the dataset with a total of 3,280 instances across the Train, Dev, and Test sets, reflecting its significant presence in legal documents. Conversely, aspects like "dịch vụ pháp lý"(legal services) and "kế toán - kiểm toán"(accounting - auditing) have relatively few samples, with only 12 and 51 instances, respectively, indicating limited data for these categories.

Additionally, the dataset is well-balanced in terms of the labels 0 (Support), 1 (Neutral), and 2 (Refute), each having approximately 3,221 instances, which ensures that the models trained on this data will be exposed to a diverse range of scenarios. The mean premise length (MPL) is consistent across the Train, Dev, and Test sets, hovering around 423 words, suggesting a uniformity in the complexity and depth of the legal texts analyzed. The Mean Hypothesis Length (MHL) across the dataset, measured in words, provides insight into the typical length of the generated hypotheses. This consistency in MHL across different splits of the dataset ensures that the complexity of the hypotheses remains stable, aiding in the uniform evaluation of NLI models.

Label Distribution: Figure 2 illustrates the label distribution in our dataset. It is evident that the labels are evenly distributed since each Premise generates 3 Hypotheses.

Aspect/Label	**Train**	**Dev**	**Test**	**Total**
bảo hiểm	66	0	6	72
bất động sản	196	24	24	244
bộ máy hành chính	2584	361	335	3280
chứng khoán	51	6	6	63
công nghệ thông tin	159	30	18	207
doanh nghiệp	54	12	6	72
dân sự	273	63	24	360
dịch vụ pháp lý	9	0	3	12
giao thông – vận tải	475	30	35	540
giáo dục	402	27	63	492
khác	608	78	96	782
kế toán – kiểm toán	27	9	15	51
lao động – tiền lương	129	15	21	165
sở hữu trí tuệ	54	12	27	93
thuế – phí – lệ phí	75	3	12	90
thương mại	63	6	6	75
thể thao – y tế	254	33	25	312
thủ tục tố tụng	130	20	9	159
tiền tệ – ngân hàng	138	6	3	147
tài chính nhà nước	576	108	48	732
tài nguyên – môi trường	144	12	21	177
vi phạm hành chính	246	15	36	297
văn hóa – xã hội	108	12	6	126
xuất nhập khẩu	54	0	12	66
xây dựng – đô thị	270	36	42	348
đầu tư	546	93	66	705
0 (Support)	2564	336	321	3221
1 (Neutral)	2564	337	322	3223
2 (Refute)	2563	338	322	3223
Total (pairs)	7691	1011	965	9667
MPL (words)	423.78	420.22	426.06	426.06
MPH (Words)	25.89	25.85	26.14	25.91

Table 1: ViLegalNLI statistics in terms of different topics, Mean Premise Length (MPL) and Mean Hypothesis Length (MHL).

Length Distribution: Figure 5 illustrates the distribution of lengths for Text, Premise, and Hypothesis, along with their respective densities (counts). The length of the Text exhibits an uneven distribution, predominantly concentrated in the range of 1,000–2,000 words, with smaller peaks observed at more distant intervals (e.g., 6,000–8,000 words). This suggests that while the texts vary in length, the majority are short to medium-length. The Premise length shows a more uniform distribution with minimal variation, peaking significantly around 400–500 words, and very few samples falling outside this range. In contrast, the Hypothesis length is strongly concentrated in the 20–30 word range, with only a few instances exceeding 40 words. This indicates that Hypotheses tend to be concise and succinct, as they are typically single sentences, whereas Premises are usually full paragraphs.

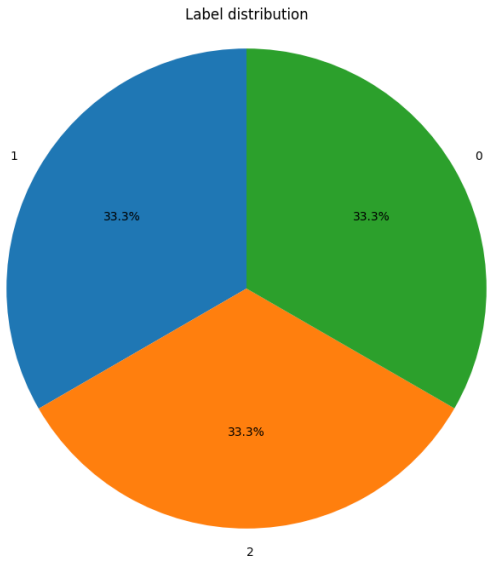

Figure 2: Label distribution

Aspect Distribution: Figure 6 illustrates the distribution of document aspects. The "Bộ máy - hành chính"dominates with over 3,000 documents, significantly outnumbering other aspects. Aspects such as "Tài chính nhà nước""Giáo dục"and "Giao thông - vận tải"have moderate representation, ranging from 500 to 1,000 documents. In contrast, aspects like "Dịch vụ pháp lý"and "Kế toán - kiểm toán"have very few documents, with only a few dozen, indicating low concentration. The disparity among aspects is striking, reflecting the uneven nature of legal data distribution across fields.

Word Frequency in Texts: Figures 3-4 illustrate the frequency of words in legal documents. Words like “quy định” (regulation), "nhân dân” (people), "ủy ban” (committee), and “thực hiện” (implementation) prominently appear with high frequency (top 4), emphasizing the focus on regulations and the activities of governmental and administrative organizations. Other phrases, such as “pháp luật” (law), “hồ sơ” (records), and “hoạt động” (activities) highlight content related to legal processes and operations. Inspired by similar studies, we used the Jaccard index to measure unordered word overlap and Longest Common Subsequence (LCS) to evaluate ordered word overlap between labels. The results are presented in the Table 2.

Jaccard Index and New Word Rate: These metrics reflect the similarity and diversity between Premise and Hypothesis. Label 0 has the highest Jaccard index (0.108) and the lowest new word rate (0.261), indicating the Hypothesis content is most similar to the Premise, with minimal use of new words. Label 1 exhibits the highest new word rate (0.387), signifying diverse and distinct content compared to other labels. Label 2 falls in the middle, with a Jaccard index of 0.082 and a new word rate of 0.293.

LCS (Longest Common Subsequence): Label 0 has the highest LCS length (34.00), demonstrating the most structural similarity between Hypothesis and Premise. Label 2 has the lowest LCS length (28.92), reflecting differences in structure or presentation.

Part-of-Speech Distribution: Nouns account for the highest proportion across all labels, especially in Label 0 (36.88%), emphasizing the focus on objects and concepts. Verbs and adjectives remain stable across groups, with Label 1 showing the highest adjective usage (1.62%), indicating a more descriptive nature. Notably, Label 2 has the highest proportion of adverbs (8.29%), suggesting detailed and emphasized explanations. In summary, label 0 demonstrates high stability and strong alignment with Premise, label 1 is diverse and novel in content. And label 2 focuses on detailed interpretations but shows less structural alignment with Premise. These findings underline the linguistic and structural differences between Hypothesis and Premise across labels, offering valuable insights for further analysis and applications.

4 Empirical Evaluation

Before entering the data into the models, we preprocess the data sets by encoding the attributes ’hypothesis’ and ’premise’ using a Vietnamese language encoder. In addition, tabular data are transformed into a structured data set format.

4.1 Baseline Models and Settings

For this dataset, we conducted experiments using four deep learning models: the BERT multilingual base model (mBERT), phoBERT, XLM-RoBERTa (XLM-R) and the cafeBERT model. For these models, we define a common set of hyperparameters as follows:

- **Learning rate:** 2e-5 *(The step size at each iteration while moving toward a minimum of a loss function)*
- **Batch size:** 32 *(Number of training samples used in one iteration to update the model’s weights.)*
- **Number of epochs:** 5 *(The total number of iterations of all the training data in one cycle for training model)*

Figure 3: Word Cloud of Test

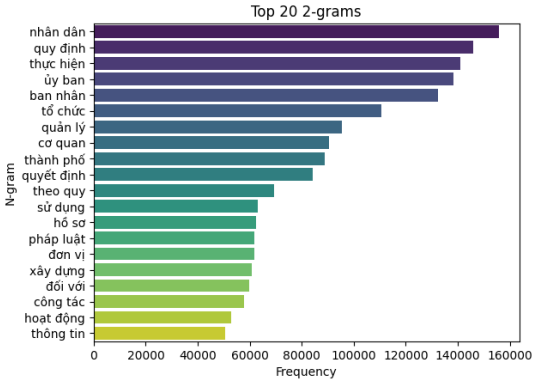

Figure 4: Top 20 ngrams

- **Weight decay:** 0.01 *(A regularization technique that helps prevent overfitting by adding a penalty to large weights.)*

4.2 Evaluation Metrics

To evaluate the experimental models, we employed several evaluation metrics, including Accuracy, F1-macro, Precision, and Recall. Our evaluation was also conducted on different aspects of the data, such as evaluating metrics based on the document domain, individual labels, and various text length intervals. This approach ensures that the evaluation is accurate, comprehensive, and objective.

4.3 Experimental Results

Based on evaluation metrics, we can conclude that most deep learning models achieve relatively high and stable performance, with the values for each metric exceeding 90 % *(table 3)*.

5 Result Analysis

Based on the experimental results *(table 3)*, it can be observed that all the evaluated models performed well and exhibited comparable performance, with each metric scoring 92 % or higher on both the development (Dev) and test sets. Among them, the XLM-RLarge model achieved the best results, with an accuracy of 94.36% and an F1-score (macro) of 94.36% on the development set, as well as an accuracy of 92.53% and an F1-score of 92.53 % on the test set.

Following closely, the CafeBERT model also delivered competitive results, with both accuracy and F1-score reaching 94.16% on the development set, and an accuracy of 92.22% along with an F1-score of 92.23% on the test set.

Lastly, the mBERT model, while achieving the lowest performance among the three models trained and evaluated on the ViLegalNLI dataset, still produced promising results, with an accuracy of 93.47% and an F1-score of 93.49% on the development set, as well as an accuracy of 92.02% and an F1-score of 92.04% on the test set.

These findings demonstrate that all three models perform effectively on the ViLegalNLI dataset. Furthermore, the results confirm the reliability of the hypothesis generation processes employed during the dataset preparation.

The prediction performance of the XLM-R model outperformed the cafeBERT model when we evaluated and analyzed results across individual domains *(tabel 4)*. However, in the fields of “xây dựng - đô thị”, “giáo dục”, “lao động - tiền lương", and “thể thao - y tế” , the cafeBERT model demonstrated superior performance. Overall, both models performed exceptionally well on the custom-collected dataset.

When evaluated by individual labels, it is evident that the XLM-R model produced higher and more stable results compared to the other models. Most metrics achieved values of 93% or higher, and the accuracy across different labels showed minimal variation. In contrast, with the cafeBERT model, the recall value for label “2” was slightly lower than that of the other labels, with recall = 89%. This result may have been influenced by the dataset during the automatic hypothesis generation process using LLMs.

The evaluation results of the deep learning models across different hypothesis length ranges in the test set are generally quite good *(Figure 7)*. Specifically, for hypothesis lengths within the ranges of (0;10] or (50;60], the accuracy and evaluation metric values are almost perfect. Additionally, the models also perform well and yield good results for hypotheses with lengths within the range of (20;40]. However, for the length range (40;50], the model’s prediction performance is not as high as for the other cases. This indicates that the length of the hypothesis also affects the model’s prediction results.

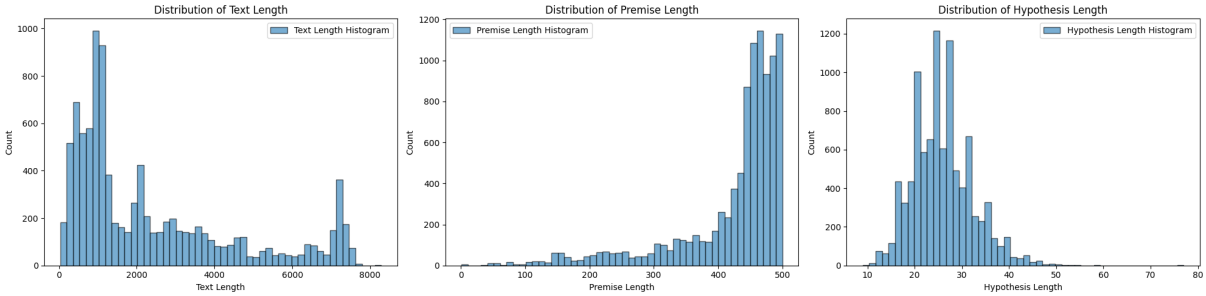

Figure 5: Distribution of text premise and hypothesis length

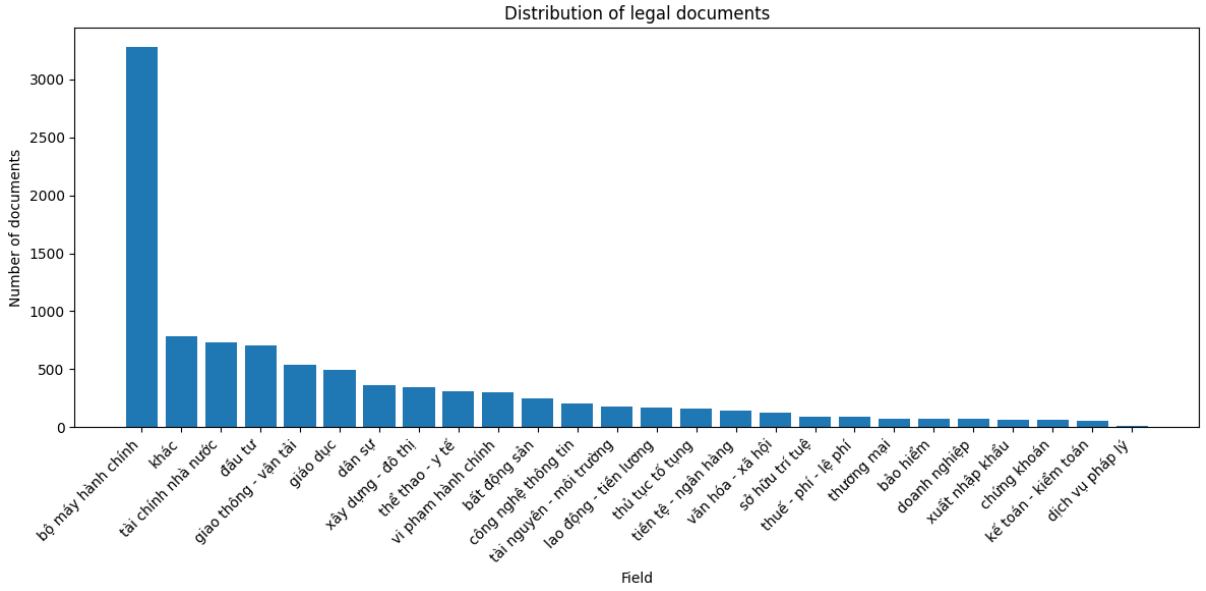

Figure 6: Distribution of aspect

Length	mBERT	XLM-RLarge	CafeBERT
(10, 20]	92	91	90
(20, 30]	92	92	93
(30, 40]	**93**	**93**	**94**
(40, 50]	92	92	92

Figure 7: Evaluation metrics per Hypothesis length (Accuracy)

Label	Jaccard	New Word Rate	LCS	Nouns	Verbs	Adjectives	Pronouns	Adverbs	Others
0	0.108432	0.260612	34.004036	36.879959	26.712625	4.241166	0.788143	3.309566	0.021528
1	0.074237	0.387026	27.660565	36.293087	24.357164	3.805218	1.615545	7.161334	0.018154
2	0.082238	0.293001	28.923984	34.723138	24.827583	3.255743	2.843682	8.291527	0.005205

Table 2: Word overlap between premise and hypothesis sentences.

Model	Dev		Test	
	Acc	F1	Acc	F1
mBERT	93.47	93.49	92.02	92.04
XLM-RLarge	**94.36**	**94.36**	**92.53**	**92.53**
CafeBERT	94.16	94.16	92.22	92.23

Table 3: Evaluation Metrics for Models on 3 Labels

Aspect	mBERT	XLM-R	cafeBERT
Bộ máy - hành chính	93	94	94
Giao thông - vận tải	94	91	94
Sở hữu trí tuệ	96	93	96
Xây dựng - đô thị	98	98	95
Tài chính nhà nước	88	85	92
Giáo dục	95	95	92
Dân sự	79	88	88
Lao động - tiền lương	86	71	90
Đầu tư	91	89	91
Thuế - phí - lệ phí	92	92	92
Kế toán - kiểm toán	87	93	93
Bất động sản	100	100	96
Vi phạm hành chính	94	94	92
Tài nguyên - môi trường	86	95	100
Công nghệ thông tin	89	83	83
Thể thao - y tế	84	80	88
Thủ tục tố tụng	78	89	78
Văn hóa - xã hội	100	100	83
Doanh nghiệp	83	83	83
Chứng khoán	100	100	83
Thương mại	100	100	100
Tiền tệ - ngân hàng	100	100	100
Khác	95	94	94

Table 4: Evaluation metrics for Different Aspects and Models.

Model	Dataset	Dev	Test
mBERT	**ViLegalNLI**	**93.47**	**92.02**
	ViNLI (Huynh et al. 2022)	67.41	64.84
XML-RLarge	**ViLegalNLI**	**94.36**	**92.53**
	ViNLI (Huynh et al. 2022)	83.02	81.36

Table 5: Evaluation Metrics for Models on Different Datasets (Accuracy)

In addition to training and evaluating the models on the ViLegalNLI dataset, we also trained and evaluated the same models on the ViNLI dataset. The overall evaluation results, using Accuracy metric, on both the dev and test sets are presented in Table 5. For the same deep learning models, our ViLegal dataset produced higher results compared to the ViNLI dataset. This further demonstrates that the ViLegalNLI dataset we collected and created is diverse and meets the required standards.

6 Error Analysis

The creation of the ViLegalNLI dataset still has some limitations as follows:

- **Error in overall model performance evaluation**: The use of test data generated by LLMs has led to unusually high model performance. This is due to the lack of diversity in the test set, which tends to be similar to the training data, potentially causing the evaluation results to not reflect the model's true generalization ability.
- **Error due to the influence of aspects on performance**: Some domains, such as Culture-Society, Commerce, and Real Estate, achieve nearly perfect performance due to the limited and less diverse data. This leads to biased evaluations, making the model appear to perform well, but in reality, it may not maintain accuracy when extended to other domains.

This evaluation helps us identify accurate directions for future development strategies.

7 Conclusion and Future Work

In this project, we introduced ViLegalNLI, high-quality corpus for evaluating Vietnamese NLI models, meet the needs of NLI research in the legal field. We constructed 9667 premise-hypothesis pairs. With hypotheses generated entirely using the LLMs model (GPT-4o) and then subjected to detailed manual screening, which is the first purely Vietnamese legal NLI corpus to date. The performance of three powerful pre-trained models (BERT Multilingual, XLM-R, CafeBERT) is all above 92% on both dev and test sets, making our feasible challenge for the Vietnamese language model in NLI tasks. We believe that ViLegalNLI will encourage the development of Vietnamese NLI research.

Leveraging state-of-the-art models on large-scale, high-quality NLI corpora (ViNLI, ViFactCheck), we hope that our corpus will accelerate progress in Vietnamese NLI research and other NLP tasks. Based on the development of Natural Language Inference (NLI) research in English, we aim to extend NLI research to Vietnamese by improving the dataset not only in terms of quantity but also quality, using more reliable data sources.

Additionally, we hope to apply our research to advanced NLP tasks in low-resource languages, such as deep semantic inference, machine reading comprehension, evidence detection and explanation generation, by applying advanced inference methods and designing functionalities for sub-tasks.

Based on the development of Natural Language Inference (NLI) research in English, we aim to extend Natural Language Inference (NLI) research to Vietnamese by enhancing the dataset in both quantity and quality. This includes generating additional data through human annotation and experiments, as well as utilizing more reliable data sources. Furthermore, we aspire to apply our findings to advanced NLP tasks in low-resource languages, such as deep semantic inference, machine reading comprehension, evidence detection, and explanation generation. This will involve leveraging advanced inference methods and designing functionalities tailored to specific subtasks.

Acknowledgments

This project is supervised by M.A. Huynh Van Tin, with foundational knowledge provided by Associate professor Nguyen Luy Thuy Ngan, PhD Nguyen Van Kiet, M.A. Nguyen Duc Vu, and other faculty members from Faculty of Information Science and Engineering, University of Information Technology, Vietnam National University HCMC.

References

[1] Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Pearson.

[2] Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 632–642. https://doi.org/10.18653/v1/D15-1075

[3] Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 1112–1122. https://doi.org/10.18653/v1/N18-1101

[4] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2475–2485. https://doi.org/10.18653/v1/D18-1269

[5] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Campagna, M. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[6] Hu, Z., Sun, M., Zhang, R., Yu, H., Xie, X., & Lu, W. (2020). OCNLI: Original Chinese Natural Language Inference Dataset. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Ham, J., Park, S., Yang, J., & Cho, K. (2020). KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[8] Mahendra, R., Ilmania, S., Wibisono, M. R. K., & Purwarianti, A. (2021). IndoNLI: A Natural Language Inference Dataset for Indonesian. Proceedings of the 2021 Workshop on Multilingual Representation Learning (MRL).

[9] Nguyen, T. H., & Nguyen, T. K. (2020). PhoBERT: Pre-trained language models for Vietnamese. Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), 3735–3745. https://doi.org/10.18653/v1/2020.coling-main.334

[10] Quyen, L. H., et al. (2022). ViNLI: A high-quality corpus for Vietnamese Natural Language Inference. Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022), 3858–3872. https://doi.org/10.18653/v1/2022.coling-main.337

[11] Sadat, W. and Caragea, C. (2022). SciNLI: Natural language inference on scientific texts. In Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 890-899.

[12] Chen, Z.; Tang, B.; and Zhou, X. (2021). NeuralLog: Bridging deep learning and logical reasoning for natural language inference. In Proceedings of the 2021 International Joint Conference on Artificial Intelligence (IJCAI), pp. 2906-2912.

[13] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 216–223. European Language Resources Association (ELRA).

[14] De Silva, S., Afsana, F., & Si, H. (2020). MED: A multilingual dataset for evaluating cross-lingual semantic similarity. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5876–5883. https://doi.org/10.18653/v1/2020.emnlp-main.475

[15] Huynh, N. D.; Pham, T. A.; and Vu, M. H. (2024). ViANLI: Adversarial Vietnamese natural language inference dataset. In Proceedings of the 2024 International Conference on Computational Linguistics and Intelligent Text Processing. (CICLing 2024).

[16] Huyen, D. T. T.; Pham, N. T.; and Le, H. V. (2024). ViHealthNLI: Vietnamese NLI dataset for healthcare domain. In Journal of Language and Knowledge Engineering, Vol. 5, Issue 2, pp. 120-135.

[17] Huynh, T. V., Nguyen, K. V., & Nguyen, N. L. T. (2022). ViNLI: A Vietnamese Corpus for Studies on Open-Domain Natural Language Inference. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).

[18] Le, H. T., To, L. T., Nguyen, M. T., & Nguyen, K. V. (2022). ViWikiFC: Fact-Checking for Vietnamese Wikipedia-Based Textual Knowledge Source. Proceedings of COLING 2022, the 29th International Conference on Computational Linguistics, 339-347.

[19] Kadiyala, R. M., Pullakhandam, S., Mehreen, K., Tippareddy, S., & Srivastava, A. (2023). Augmenting Legal Decision Support Systems with LLM-based NLI for Analyzing Social Media Evidence. Proceedings of the 2023 Conference on Legal Informatics and Technology. University of Maryland, University of Wisconsin, Traversaal.ai, University of South Florida.

[20] Koreeda, Y., & Manning, C. (2023). Con-

tractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).