ViFC: A New Benchmark For Evaluating Machine Reasoning For Vietnamese Fact-Checking

Hung Tuan Le [1,2], **Manh Trong Nguyen** [1,2], **Long Truong To** [1,2], **Tri Thien Nguyen**[1,2], **Kiet Van Nguyen** [1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{21520250, 21520343, 21521101, 21522707}@gm.uit.edu.vn
kietvn@uit.edu.vn

Abstract

The fact-checking task aims to verify the truthfulness of a statement based on available knowledge. However, current research on this task needs to focus on challenging the reasoning abilities of models. Recognizing this limitation, we developed a dataset with manually crafted claim and explanation pairs, drawing data from Wikipedia. Additionally, we analyzed various linguistic aspects within the dataset, such as New POS Tagging, New Word Rate, and Word Overlapping. We experimented with state-of-the-art models on the two tasks: veracity prediction and explanation generation. The best results achieved by the models (PhoBERT$_{Large}$ performing 40.86% in F1-score for prediction, and BARTpho - Word reaching 19.32% in BLEU score for generation) fall short of human performance. These results demonstrate that our corpus challenges machine reasoning abilities in Vietnamese fact-checking.

1 Introduction

Fact-checking is a task that aims at verifying information based on existing knowledge. This verification task is typically undertaken by a group of journalists, for example, PolityFact[1] and FactCheck[2] which play an essential role in online Fact Verification, as misinformation is prevalent on news websites. However, manual fact-checking is time-consuming because fact-checkers often need to reach out to many potential sources for evidence. This approach becomes insufficient given the rapid pace of information updates on social media. Automatic fact-checking has emerged, intended to quickly detect and verify inaccurate information on social media.

Fact-Checking is a complex task that requires the extraction and reasoning of multiple pieces of evidence. In a survey on fact-checking by Guo et al., fact-checking is divided into four tasks: Claim Detection, Evidence Retrieval, Verdict Prediction, and Justification Production. Among these, Verdict Prediction and Justification Production are two tasks that demand the model to infer and provide well-reasoned judgments along with explanations for the judgments based on the evidence extracted in the Evidence Retrieval task.

In recent years, the fact-checking problem has gained attention from the research community, leading to the creation of various datasets constructed based on information from Wikipedia[3], ranging from datasets with complex claims challenging the reasoning abilities of models like HOVER (Jiang et al., 2020), to those like FEVEROUS (Aly et al., 2021). Alongside the emergence of these complex datasets are methods aimed at enhancing models' reasoning and explanation capabilities, such as the use of Graph-based (Liu et al., 2020b; Nguyen et al., 2020) and joint (Atanasova et al., 2020). However, research has primarily focused on English or multilingual languages, leaving studies on Vietnamese with limitations in both data sources and methodologies.

To contribute to developing machine reasoning in Vietnamese NLP research, especially in Fact-Checking, we introduce a new open-domain benchmark Corpus designed for evaluating Vietnamese machine reasoning: ViFC - Vietnamese Fact-Checking on Wikipedia. The benchmark includes 1000 sample and is manually constructed based on evidence extracted from Wikipedia pages. Taking advantage of FEVER (Thorne et al., 2018) labels, our corpus comprises three label classes: SUPPORT, REFUTE, and NOTENOUGHINFORMATION (NEI). In our dataset, given a context paragraph, annotators must extract evidence sentences from the context. Subsequently, they rewrite the claim sentence based on the labels derived

[1] https://www.politifact.com/
[2] https://www.factcheck.org/
[3] https://vi.wikipedia.org/

from the information in the extracted evidence sentences. Additionally, to challenge the model's inference and language generation abilities, for each evidence-claim-label set, annotators provide explanations for why the claim is labeled as SUPPORTS, REFUTES, or NEI based on the information found in the evidence.

Before corpus construction, we developed a suitable labeling guideline and a user-friendly labeling tool. Annotators, who are Vietnamese native speakers with an education background beyond high school, undergo training and inter-rater agreement assessments to ensure the quality of the corpus.

We designed two experiments to evaluate the model's inference capabilities: verdict prediction and justification production. In the verdict prediction task, the model must infer from the provided information in the evidence sentences to verify the claim. We evaluated the models using state-of-the-art pre-trained language models that have shown significant results in Vietnamese NLP tasks, including multilingual models such as Multilingual BERT (mBERT) (Kenton and Toutanova, 2019), XLM-RoBERTa (XLM-R) Conneau et al. (2020), and Info-XLM (Chi et al., 2021), as well as monolingual models designed for Vietnamese, such as PhoBERT (Nguyen and Tuan Nguyen, 2020), ViBERT (Tran et al., 2020), and ViT5 (Phan et al., 2022).

The model takes an evidence-claim-label set as input for the justification production task and produces an explanatory sentence. In this task, we employed SOTA pre-trained sequence-to-sequence language models, including multilingual models like Multilingual BART (mBART) (Liu et al., 2020a) and Multilingual T5 (mT5) (Xue et al., 2021), and also monolingual models for Vietnamese, such as ViT5 (Phan et al., 2022) and BARTpho (Tran et al., 2022).

The contributions in this paper are outlined as follows:

- Firstly, we introduce ViFC, a new benchmark dataset for evaluating machine reasoning in Vietnamese fact-checking. This dataset includes a comprehensive 1000 set of evidence-claim-explanation triplets, meticulously constructed from over 212 Wikipedia articles by a team of annotators with an inter-rater agreement reaching 90.75% Fleiss' κ-agreement.

- Secondly, we design two experiments to assess the reasoning capabilities of models, encompassing verdict prediction and justification production, utilizing state-of-the-art pretrained transformer-based models.

- Next, we analyze the corpus from various linguistic aspects to gain more insights into language features in constructing claim and explanation sentences.

- Finally, we scrutinize the experimental results of the models across different aspects of the corpus to highlight the challenges in reasoning that the corpus presents.

2 Related Work

2.1 Fact-checking verdict prediction

Verdict prediction in fact-checking is a task that requires the model to infer based on the information provided in the evidence to authenticate the information in the evidence (Guo et al., 2022). In recent years, several datasets have been published to facilitate the evaluation and development of methods to improve models' reasoning capabilities. These datasets include manually curated datasets for research purposes such as FEVER (Thorne et al., 2018), VitaminC (Schuster et al., 2021), TabFact (Chen et al., 2020), as well as datasets with naturally occurring claim sentences on topics such as politics and society, for example, PolitiFact (Vlachos and Riedel, 2014) and Liar (Wang, 2017) extracted from fact-check websites, FakeCovid (Shahi and Nandini, 2020) removed from Covid-19 news websites, or PUBHEALTH (Kotonya and Toni, 2020) collected from fact-check websites in the public health domain.

Notably, these datasets still do not pose significant challenges for models, as claim sentences can often be inferred from just a few pieces of information in a typical evidence sentence (Jiang et al., 2020), as seen in the FEVER dataset (Thorne et al., 2018) where claim sentences constructed from multiple evidences only constitute 16.82% of the total dataset.

To bridge this gap, some datasets have been introduced with claim sentences constructed from multiple evidence sentences, such as FEVEROUS (Aly et al., 2021) and HOVER (Jiang et al., 2020). Alongside these complex datasets, various methods have been proposed to address the challenges posed by these datasets, such as applying Graph-based models (Zhou et al., 2019; Zhong et al., 2020;

Nguyen et al., 2020; Barnabò et al., 2023) to facilitate inference across multiple pieces of information. Recently, Pan et al. introduced a framework that decomposes complex claims into simpler sub-tasks, making inference more manageable.

However, most of this research is primarily conducted in English, leaving a gap in the availability of complex and challenging datasets for the Vietnamese language. To address these issues, we propose a new benchmark for evaluating machine reasoning in Vietnamese fact-checking, with claims manually constructed from various pieces of evidence.

2.2 Fact-Checking justification production

The task of justification production has recently gained attention from the research community, as in practice, merely stating the veracity of the claim still needs to be more persuasive for the audience. This task challenges the model's ability to infer and generate language when providing an explanation for the reasons behind the verdict about the claim based on evidence.

Currently, explanations in existing datasets can be categorized into two types: those extracted from readily available sources, such as the LIAR-PLUS dataset (Alhindi et al., 2018)—an extended version of the LIAR dataset (Wang, 2017), supplemented with explanations extracted from comments on fact-checking websites and the FactEx (Althabiti et al., 2023) dataset with explanations removed from various fact-checking websites, including SpanEx (Choudhury et al., 2023) with phrase-level explanations directly extracted from evidence used for verification. Besides the methods mentioned above of constructing explanations, there are explanations generated by humans based on datasets such as PUBHEALTH (Kotonya and Toni, 2020) (a dataset in the public health domain) and ChartCheck (Akhtar et al., 2023) (a dataset for fact-checking on charts). However, in the two methods of constructing claim sentences, the human-generated explanation method is considered more challenging than the remaining methods. It requires the model to infer and present arguments from the information provided so that the explanation aligns closely with human-generated explanations (Celikyilmaz et al., 2020).

With the emergence of these datasets, solution methods for the problem have also been rapidly developed, such as joint models for both verdict prediction and justification production (Atanasova et al., 2020), utilizing attention weights to select supporting information from the evidence (Popat et al., 2017; Cui et al., 2019; Lu and Li, 2020), or summarizing information in the evidence related to the claim to form an explanation (Kotonya and Toni, 2020; Atanasova et al., 2020; Jolly et al., 2022).

In our dataset, annotators write an explanation based on their reasoning for each evidence-claim-label set, relying only on the information in the evidence. Additionally, we provide evaluation criteria for writing explanations to ensure that the explanations exhibit high inferential and clear reasoning, thereby challenging the model's inference and generation capabilities.

3 Corpus Creation

Our corpus was developed through 4 phases (see Figure 1), encompassing data collection (see section 3.1), annotator training (see section 3.2), corpus generation (see section 3.3), and corpus validation (see section 3.4). In addition, in order to gain a deeper understanding of the linguistic features of the corpus, we conducted a corpus analysis across various linguistic aspects (see section 3.5).

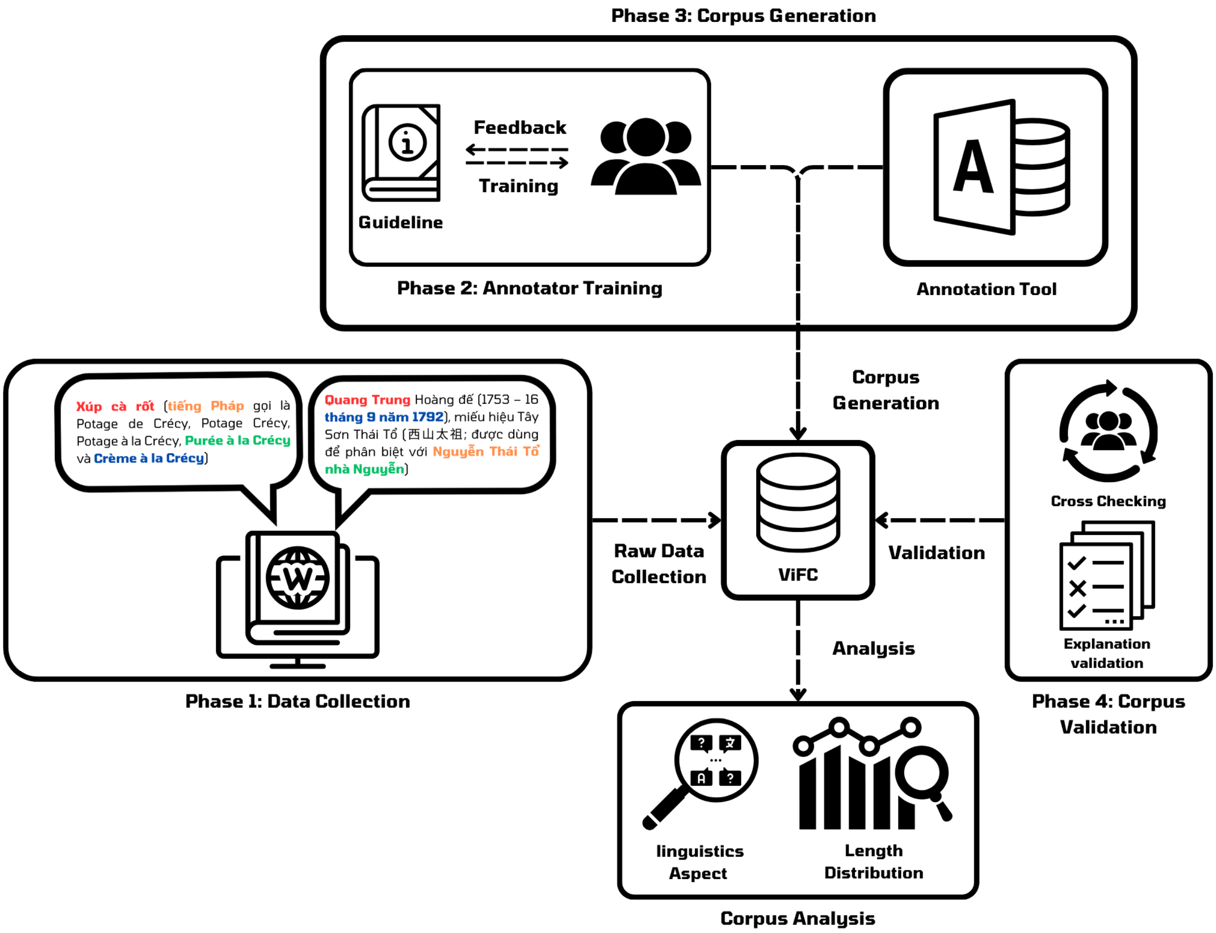
3.1 Data Collection

We chose Wikipedia[4] as the primary data source for building the ViFC corpus. Wikipedia is an open-source encyclopedia with millions of articles covering various topics, ranging from everyday life, society, and health to natural sciences such as physics, mathematics, and chemistry. With such a diverse data source, Wikipedia provides a rich knowledge base. In this study, we do not assess the authenticity of the data but rather emphasize the data coverage. Wikipedia satisfies the coverage requirements by encompassing a wide range of topics across different domains of life. We extracted 2895 paragraphs from 212 Wikipedia articles covering diverse subjects such as health, history, geography, and science.

3.2 Annotator Training

Annotators who are native speakers of Vietnamese with education backgrounds beyond high school are selected. All annotators undergo a training process to familiarize themselves with the data construction process before entering the official annotation process. We use Fleiss κ score to assess the agreement

[4] https://vi.wikipedia.org/


Phase 3: Corpus Generation
Feedback
Training
Guideline
Phase 2: Annotator Training
A
Annotation Tool
Xúp cà rốt (tiếng Pháp gọi là Potage de Crécy, Potage Crécy, Potage à la Crécy, Purée à la Crécy và Crème à la Crécy)
Quang Trung Hoàng đế (1753 – 16 tháng 9 năm 1792), miếu hiệu Tây Sơn Thái Tổ (西山太祖; được dùng để phân biệt với Nguyễn Thái Tổ nhà Nguyễn)
W
Phase 1: Data Collection
Raw Data Collection
Corpus Generation
ViFC
Validation
Analysis
Cross Checking
Explanation validation
Phase 4: Corpus Validation
linguistics Aspect
Length Distribution
Corpus Analysis

Hình 1: ViFC Corpus Creation

among annotators.

The first annotator is provided with guidelines on data construction, including how to select evidence, rewrite claim sentences for three labels (SUPPORT, REFUTE, and NEI), and finally, write explanations for each claim. Annotators are tested on a set of 50 data samples, and based on this dataset, the agreement among annotators is evaluated for the process of relabeling evidence-claim-explanation sets. If the agreement among annotators is above 0.9, annotators will officially proceed to the corpus construction process.

3.3 Corpus Generation

First, annotators are provided with a paragraph from a Wikipedia article. Subsequently, annotators read and write a claim based on the information provided in the context of three labels: SUPPORT, REFUTE, and NOT ENOUGH INFORMATION (NEI). The labels are defined as follows:

- **SUPPORT**: Annotators create a claim that we can determine to be accurate based only on information from the evidence.

- **REFUTE**: Annotators generate a claim that we can evaluate as false based only on the information in evidence.

- **NOT ENOUGH INFORMATION (NEI)**: Annotators create a claim that we cannot choose to be true or false based only on the evidence provided.

After writing the claim, annotators must select sentences from the context used to construct the claim. While selecting evidence, annotators must answer questions such as "Is the chosen evidence related to the information in the claim?""Does the chosen evidence support the inference in the claim?"and "Do we have enough evidence to verify the information in the claim with the chosen evidence?"Finally, annotators have to write an explanation corresponding to each evidence-claim-label set. Similar to the claim-writing process, annotators answer questions to ensure the quality of constructing explanations, such as "Does the explanation correctly explain the information that influences labeling the claim?""Does the explanation provide enough information to influence labeling the claim?"and "Does the explanation use external knowledge beyond the provided evidence?"Throughout the corpus construction process, annotators are prohibited from using external knowledge beyond the provided evidence.

For each provided paragraph, annotators must generate evidence-claim-explanations for the three labels. With data construction methods like these, we increase diversity in semantics and expression on the same set of information.

Our team identified difficulties in building claims and explanations throughout the data construction process. Annotators faced challenges when writing claims under the SUPPORT label, especially when information in the selected evidence was disparate, requiring annotators to infer more to find new information without using external information. When writing explanations, annotators encountered difficulties in ensuring logical reasoning and avoiding the inclusion of information beyond the scope of evidence.

3.4 Corpus Validation

In the corpus validation section, we evaluate the dataset on two tasks: cross-validation (see section 3.4.1) and explanation validation (see section 3.4.2).

3.4.1 Cross Validation

We selected 50 labeled data samples, and then the labels were removed. For this unlabeled dataset, annotators were tasked with re-labeling each evidence-claim pair. In cases where fewer than 2 annotators agreed on a label, the pair of claim sentences would be marked and removed from the corpus. The statistical results showed that, out of the 50 data samples, there were no instances where fewer than 2 annotators agreed on a label. Additionally, we calculated inter-annotator agreement using the Fleiss κ score, and the obtained result for the task was 90.75%.

3.4.2 Explanation Validation

To evaluate the process of constructing explanations, we will randomly select 50 data samples, each consisting of one set of evidence-claim-label-explanation. For each data sample, annotators are required to assess the explanations based on tow criteria as follows: "Does the explanation provide enough information that directly influences labeling the claim?"and "Does the explanation correctly explain the information that directly influences labeling the claim?". For each criterion the explanation achieves, annotators will mark it as "1"; otherwise, keep it as "0."The final result will be evaluated through Fleiss κ, and the statistical result indicates a consensus among annotators at 90.31%.

3.5 Corpus Analysis

3.5.1 Overall Statistics

Our dataset consists of 1000 data points and has been randomly divided into three sets: 80% for the training set (Train), 10% for the development set (Dev), and 10% for the test set. The distribution of each label across the three datasets is described in Table 1.

	Train	Dev	Test
SUPPORT	261	35	38
REFUTE	267	37	30
NEI	272	28	32

Bảng 1: Overall statistics of ViFC corpus

3.5.2 Length Distribution

To analyze the trends in writing claim and explanation sentences in the corpus, we measured the length of both claim and explanation sentences and the evidence. The length distribution is presented in figure 2.

The lengths of claim and explanation sentences are concentrated mainly in the 0 to 60 words range, with sentences longer than 60 words forming a small portion of the corpus. Notably, in the range of 0 to 60 words, there is significant similarity in the distribution between claim and explanation sentences. This indicates that explanations avoid mere repetition of information from evidence or claim and elaborate, reason, and infer key insights crucial for predicting the claim's veracity. The length distribution of evidence differs from that of claim and explanation, with evidence lengths ranging from 20 to over 150 words and a notable concentration in the 40 to 100-word range. This is understandable as annotators need to extract sufficient information from the evidence, considering the substantial information provided.

3.5.3 New Word Rate

In our corpus quality assessment, we meticulously examine lexical diversity by quantifying the emer-

		Overlapping		New Word Rate (%)	New Pos Tagging (%)					
		Jaccard	LCS		Noun	Verb	Adjective	Preposition	Adjunct	Other
Claim	SUP	24.88	78.99	26.35	32.61	26.39	8.63	8.52	7.02	16.83
	REF	23.78	72.64	27.88	34.83	25.66	7.91	9.02	6.28	16.30
	NEI	25.11	81.08	30.29	34.35	24.78	8.65	8.12	6.71	17.39
Explanation	SUP	49.92	117.56	24.81	39.69	25.15	6.65	6.00	4.10	18.41
	REF	38.66	71.31	26.64	31.13	25.27	8.98	4.73	9.98	19.90
	NEI	22.18	39.16	44.19	26.78	33.22	5.38	4.95	12.11	17.57

Bảng 2: Corpus analysis in terms of linguistic aspects

gence of novel words within claim assertions that are absent in evidentiary text and unique lexical units present in explanatory discourse yet to be found in claims or evidence. Utilizing the VnCoreNLP toolkit (Vu et al., 2018) for word segmentation, we observe a discernibly lower incidence of new word introduction in claims associated with SUPPORT and REFUTE labels, with the SUPPORT category registering the minimal novel word incidence among the labels examined (Table 2). In stark contrast, claims tagged with the NEI (Not Enough Information) label exhibited a markedly higher rate of new lexical entries, a trend congruent with the NEI label's inherent requirement for referencing information beyond the provided evidence.

Furthermore, a parallel trend is identified in the construction of explanatory sentences: the NEI label consistently manifested the highest rate of new word utilization, followed by REFUTE and SUPPORT labels. This pattern persist, notwithstanding the expansive lexicon encompassing evidence and claims, with the rate of new words consistently exceeding 24%. This phenomenon suggests that annotators strategically employ a broader vocabulary to articulate inferences, thereby enriching the explanations without extending beyond the parameters of the given evidence-claim nexus. Complementing this linguistic analysis, a Part-Of-Speech (POS) scrutiny is conducted on the newly introduced words to elucidate the annotators' tendencies in corpus construction. By deploying VnCoreNLP (Vu et al., 2018) for POS tagging, we discern a preponderance of Nouns and Verbs, indicating that these grammatical categories are pivotal in the formulation of both claims and explanations, reflecting the annotators' linguistic patterns in the dataset development process (Table 2).

3.5.4 Word Overlapping

In the study by McCoy et al., word overlapping influenced the model's inference capabilities. Therefore, we conducted word overlapping calculations on the corpus using two parameters: Jaccard for unordered word overlap rate and LCS index (Longest Common Sub-sequence) for ordered word overlap rate. We analyzed claims and explanations using the method described in Section 3.5.3. The statistical results are presented in Table 2.

The overlapping rate across all three labels showed no significant difference with the Jaccard index for claims. However, with the LCS index, there was a pronounced difference, with the NEI label achieving the highest value, followed by SUP, and finally, REF. The highest value for the NEI label suggests that annotators utilize more information from the evidence and then extend their knowledge to construct the claim.

In explanations, both Jaccard and LCS indices indicated that the SUPPORT label had the highest rate and index among the three labels, followed by REF and NEI. This surprising phenomenon suggests that for explanations in the SUPPORT label, annotators use more information in the evidence and claim to form a robust argument that helps explain the inferred information in the claim, contributing to creating a high-quality explanation.

4 Baseline Models

4.1 Verdict Prediction

In Natural Language Processing (NLP), leveraging large-scale language models grounded in the transformer-based architecture has become a pivotal approach, particularly when handling constrained datasets. Our methodology incorporates an array of pre-eminent multilingual and monolingual pre-trained language models at the NLP field's vanguard. These models encompass the likes of Multilingual BERT (mBERT) (Kenton and Toutanova, 2019), XLM-RoBERTa (XLM-R) (Conneau et al., 2020), and InfoXLM (Chi et al., 2021) — which are trained in a multitude of languages exceeding ninety vernaculars, inclusive of Vietnamese.

Furthermore, we harness the prowess of monolingual models explicitly crafted for the Vietnamese lexicon, such as PhoBERT (Nguyen and Tuan Nguyen, 2020), ViBERT (Tran et al., 2020), and ViT5 (Phan et al., 2022). ViT5 is predicated on an encoder-decoder transformer architecture and is fortified by the T5 self-supervised pre-training framework. PhoBERT emerges as a paragon, trained extensively on a corpus comprising 20GB of Vietnamese Wikipedia and news text, boasting 135 million parameters in its base iteration and an augmented 370 million parameters for the large variant. In its most recent iteration, PhoBERT$_{base-v2}$, the model has been refined on a formidable 120GB of Vietnamese text derived from the OSCAR-2301 dataset[5].

Given the intrinsic nature of the Vietnamese language, wherein a single word may be comprised of an array of syllables, our empirical evaluation spans two experimental paradigms. One utilizes word-based inputs employing models such as PhoBERT and ViBERT in conjunction with the VNCoreNLP toolkit (Vu et al., 2018) for meticulous word segmentation. The other paradigm adopts a syllable-based approach, using models such as mBERT, XLM-R, and InfoXLM, alongside ViT5, to delve into the syntactic and semantic intricacies of the Vietnamese language. This bifurcated approach enables a comprehensive analysis of the language's nuanced characteristics, cementing the efficacy of the deployed models across various linguistic dimensions.

4.2 Justification Production

The task of Justification Production is a pivotal component in natural language processing, particularly in the realm of fact-checking, where the generation of coherent, relevant, and plausible explanations for claims is crucial. We choose state of the art encoder - decoder transformer based model for Vietnamese such as ViT5 (Phan et al., 2022) (both Base and Large versions) and BARTpho (Tran et al., 2022) (in its four variants: Word, Word Base, Syllable, and Syllable Base). Monolingual models, tailored to the intricacies of the Vietnamese language, exhibit a remarkable ability to generate justifications that are lexically and syntactically aligned with the nuanced demands of the language.

On the other hand, the multilingual models, we use mT5 (Xue et al., 2021) (xsmall and base versions) and mBART$_{Large-50}$ (Liu et al., 2020a), to bring to the table their extensive cross-lingual capabilities. Although not exclusively trained on Vietnamese, these models leverage their diverse linguistic training to offer a broad understanding of various linguistic patterns and structures, which can be beneficial when dealing with cross-border information or multilingual datasets. The mBART$_{Large-50}$, with its expansive training corpus encompassing more than 50 languages, stands out for its ability to grasp a wide array of linguistic nuances, making it a versatile tool for justification generation across languages.

4.3 Evaluation Metrics

For the task of verdict prediction, our evaluation framework centers on the deployment of accuracy and the F1-score as the primary metrics. Accuracy is the most direct measure, reflecting the proportion of verdicts correctly predicted by the model out of the total cases, thus providing a straightforward assessment of performance. The F1-score complements this by balancing the precision and recall, offering a more nuanced view that accounts for the model's ability to handle class imbalance effectively.

For justification production task, we adopt a suite of metrics, including Rouge-1, Rouge-2, Rouge-L, BLEU, and BERTscore. Rouge-1 and Rouge-2 focus on the overlap of unigrams and bigrams between the generated justifications and a set of reference texts, providing insights into the lexical precision of the model's output. Rouge-L considers the longest common subsequence, offering a perspective on the generated text's structural coherence. BLEU, traditionally used in machine translation, evaluates the grammatical and semantic accuracy of the generated justifications against multiple references. BERTscore, leveraging contextual embeddings, assesses the semantic fidelity of the justifications, ensuring that the model's outputs align closely with the intent and meaning of the ground truth. This array of metrics ensures a multifaceted assessment, capturing the quality of generated justifications across various linguistic dimensions.

5 Experiment Result

5.1 Verdict Prediction

Upon reviewing the results presented in table 3, it is evident that verdict prediction task has yielded a diverse range of outcomes across both word-based

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

	Model	Dev		Test	
		Acc	F_1	Acc	F_1
Word	ViBERT	34.67	34.21	19.33	19.15
	PhoBERT$_{Base}$	46.00	46.06	35.33	34.11
	PhoBERT$_{Base-v2}$	46.67	46.47	38.00	36.85
	PhoBERT$_{Large}$	**51.33**	**49.73**	**42.67**	**40.86**
Syllable	mBERT	36.00	35.01	36.00	31.29
	ViT5$_{Base}$	38.00	38.02	32.00	31.44
	ViT5$_{Large}$	40.67	39.58	32.51	31.79
	XLM-R$_{Base}$	39.33	39.24	28.00	21.33
	XLM-R$_{Large}$	**42.00**	38.60	**38.67**	**32.12**
	InfoXLM$_{Base}$	30.67	28.17	29.33	26.71
	InfoXLM$_{Large}$	41.33	**41.08**	31.33	28.57

Bảng 3: Veracity prediction result on Dev and Test.

and syllable-based pre-trained language models. The word-based models, with a particular emphasis on the PhoBERT series, showcase a robust performance, with PhoBERT$_{Large}$ leading the pack. This superior performance is likely attributable to the model's ability to capture the intricate nuances of the Vietnamese at the word level. PhoBERT$_{Base-v2}$ also demonstrates commendable results, hinting at the beneficial impact of training on an expanded dataset.

In contrast, syllable-based models like mBERT, ViT5, XLM-R, and InfoXLM, while generally effective, do not reach the high performance levels of the word-based counterparts. The large versions exhibit a tendency to outshine the base versions in the development set, suggesting that the additional parameters offer an edge in learning the task's complexities. However, this does not uniformly translate to the test set, where overfitting or data distribution disparities may have influenced the outcomes. Notably, XLM-R and InfoXLM models, both in their base and large configurations, deliver competitive results, yet they show a drop in test performance compared to development scores. This pattern could imply the models' enhanced capacity for feature learning, yet a susceptibility to overfitting.

5.2 Justification Production

Examining the results in table 4 for justification production task, it is discernible that the monolingual models outperform their multilingual counterparts, across most metrics on both development and test datasets. Focusing on the monolingual models, large versions typically surpass base versions, indicating that the increased model complexity can indeed translate to better performance in generation task. We can see ViT5$_{Large}$ archieves a higher Rouge 1 and Rouge 2 score compared to ViT5$_{Base}$, which suggests that it is more adept at capturing both unigram and bigram overlaps with reference text. This trend is further substantiated by its superior Rouge-L score, implying enhanced sentence-level coherence. BARTpho, with its word-based variant, register with the highest BLEU score among all models on the development set, underscoring its proficiency in generating sentences that are closer to human reference sentences. Moreover, this model variant also records the highest BERTScore, reflecting its ability to produce semantically relevant and contextually rich language outputs.

Conversely, the multilingual models, namely mBART$_{Large-50}$ and mT5, display relatively weaker performance, particularly mT5$_{Small}$ and mT5$_{Base}$. This could be attributed to the inherent challenges posed by multiple languages, which may dilute the model's capacity to fine-tune its generative capabilities for any single language.

6 Result Analysis

To gain more insight how the features of dataset affect models, we select two best 'word' and 'syllable' based models. We use PhoBERT$_{Large}$ and XLM-R$_{Large}$ to analyze the impacts of evidence quantity and claim sentence length on verdict prediction task.

Firstly, we analyze how the amount of evidence impacts the accuracy of label predictions in language models. We investigate whether the size of the evidence corpus affects the models' ability to accurately predict outcomes. This study aims to understand how information density influences model training and verdict prediction. Following this, we examine the effect of claim length on these predictions. We propose that longer claims, with their inherent complexity, may push the models to their interpretative limits, affecting their prediction accuracy. Our goal is to identify the claim length at which model performance is optimized, aiding in the selection and tuning of models for effective fact-checking.

6.1 Verdict Prediction

The figure 3 illustrates a striking trend in the correlation between the quantity of evidence and the prediction accuracy of two language models, XLM-R$_{Large}$ and PhoBERT$_{Large}$. As the number of evi-

Model	Dev					Test				
	Rouge 1	Rouge 2	Rouge L	BLEU	BertScore	Rouge 1	Rouge 2	Rouge L	BLEU	BertScore
ViT5$_{Base}$	49.45	31	39.27	9.1	76.22	54.58	35.57	43.45	12.92	77.31
ViT5$_{Large}$	**52.03**	**35.62**	**42.48**	10.94	77.18	50.21	31.95	39.38	9.05	75.57
BARTpho$_{Base}$ - Syllable	49.41	32.32	39.98	7.65	76.2	47.51	28.94	38.3	6.66	75.85
BARTpho$_{Base}$ - Word	53.64	34.15	42.43	**14.88**	**77.96**	55.99	34.67	44.11	18.22	78.64
BARTpho - Syllable	47.86	30.66	38.86	6.87	76.26	51.2	32.32	40.3	8.59	76.51
BARTpho - Word	51.89	31.93	40.45	13.33	77.66	**65.64**	**36.06**	**44.19**	**19.32**	**78.71**
mBART$_{Large}$	45.4	16.04	30.16	10.07	68.31	46.97	16.52	30.21	8.49	68.49
mT5$_{Base}$	14.88	4.78	13.27	0.61	59.8	13.73	5.64	12.24	0.47	60.32
mT5$_{Small}$	9.08	2.94	7.7	0.58	57.2	8.2	2.32	6.65	0.29	56.78

Bảng 4: Explanation generation result on Dev and Test

dence items increases, there is a pronounced decline in the proportion of correct predictions for both models. This suggests a potential complexity threshold within these models, where an overabundance of information may paradoxically obfuscate rather than elucidate the veracity of the prediction.

For the XLM-R$_{Large}$ model, the prediction accuracy peaks when only a single piece of evidence is presented and deteriorates as more evidence is introduced. This degradation could be due to the model's difficulty in synthesizing multiple strands of information into a coherent narrative essential for accurate verdict prediction. PhoBERT$_{Large}$ exhibits a similar trend, although the decline in accuracy is not as steep compared to XLM-R. This model maintains a more consistent performance across varying evidence quantities, which may be attributed to its specific optimization for the Vietnamese language and potentially better handling of the nuances.

Figure 4 show the analysis of the impact of claim length on XLM-R and PhoBERT models within a fact-checking context reveals distinct trends. XLM-R shows decreasing accuracy as claim length increases, suggesting it may be more suited for shorter, less complex claims. In contrast, PhoBERT demonstrates increased robustness with longer claims, possibly due to its better grasp of Vietnamese linguistic subtleties. Both models perform well with moderate-length claims, but XLM-R's performance drops with longer claims, unlike PhoBERT, which remains relatively stable. These observations highlight PhoBERT's consistency across claim lengths and suggest that XLM-R may benefit from optimizing claim lengths during preprocessing for improved performance in fact-checking applications.

6.2 Justification Production

We analyzed the generation results on the development set of the model that achieved the best performance on the test set, namely BARTPho - Word. For each development set, we compiled statistics on the generated sentences based on four criteria, namely:

- **Sufficient Information**: The explanation provides enough information that influences the prediction of the claim's veracity.
- **Correct Information**: The explanation provides correct information that influences the prediction of the claim's veracity.
- **Logical**: The information in the claim is interpreted logically, adhering to sound reasoning without conflicts among details.
- **Smooth**: The wording and information in the explanation are clear and do not pose difficulties in comprehension.

We will read it, and for each criterion, if the generated sentence meets the criterion, we will mark it as ’1’; otherwise, we will keep it as ’0’. The results are summarized in Table 5. Most generated explanations achieve a smooth flow, but in other aspects, such as ’Sufficient Information’ and ’Correct Information,’ significant results have not been obtained. This indicates that the model still struggles to recognize informative patterns influencing the prediction of claim veracity. Regarding the ’Logical’ criterion, even though the statistical result reaches 54%, in cases of correct logic, the model often directly extracts one or more pieces of evidence without providing meaningful inference between crucial pieces of information. For cases with incorrect logic, the model generates sentences with contradictions within the same information sample. About 13% of the sentences satisfy all four criteria. However, these cases mainly involve a mix of SUPPORT and NEI labels, where the generated explanations extract essential information but fall short of providing inferences beyond the evidence.

From the above analyses, it is evident that on our dataset, the models are still unable to achieve explanations comparable to those in our constructed dataset. However, this also proves that ViFC challenges the inference and language generation capabilities of Vietnamese sequence-to-sequence models.

	SuffIn	CorrIn	Logical	Smooth	Total 4 Criteria
Ratio (%)	24	37	54	91	13

Bảng 5: The ratio of generated sentences meeting each criterion and all four criteria.

7 Conclusions and Future Directions

In our recent study, our team developed the ViFC dataset—a comprehensive, multi-domain benchmark specifically crafted to assess the reasoning and generative capabilities of machines in the Vietnamese fact-checking domain. This dataset encompasses 1000 cases, each containing a claim based on evidence and accompanied by a detailed, standard explanation created by humans. Through linguistic analysis, we extracted significant insights related to the dataset construction model, insights that hold considerable potential to support the development of dataset creation methods, within the scope of the Vietnamese language.

The experimental outcomes on the ViFC dataset, particularly in the areas of decision processing and explanatory synthesis, were optimized by PhoBERT$_{Large}$ and BARTpho - Word. However, these models only achieved a maximum F1-score of 40.86% and a BLEU score of 19.32%, highlighting the significant challenges that Vietnamese pre-trained language models currently face in terms of reasoning and creative capabilities in the fact-checking field.

The results emphasize the nascent reasoning and language generation abilities of current Vietnamese pre-trained language models, which have yet to reach an adequate threshold. The primary cause identified for this shortfall is the dataset's limited size. Therefore, our team's future research direction is to develop and expand the corpus in terms of quality and quantity by integrating a broader range of diverse data sources. Moreover, we plan to apply modern methods, such as graph-based models or prompt tuning on large language models like GPT-4 and Mixtral, to enhance the reasoning and language generation abilities of the models.

References

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2023. Chartcheck: An evidence-based fact-checking dataset over real-world chart images. *arXiv preprint arXiv:2311.07453*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2023. Generative ai for explainable automated fact checking on the factex: A new benchmark dataset. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 1–13. Springer.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Sagnik Choudhury, Pepa Atanasova, and Isabelle Augenstein. 2023. Explaining interactions between text spans. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12709–12730, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2961–2964, New York, NY, USA. Association for Computing Machinery.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10):500.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020b. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1165–1174, New York, NY, USA. Association for Computing Machinery.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 624–643, Online. Association for Computational Linguistics.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. 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*.

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

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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

William Yang Wang. 2017. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

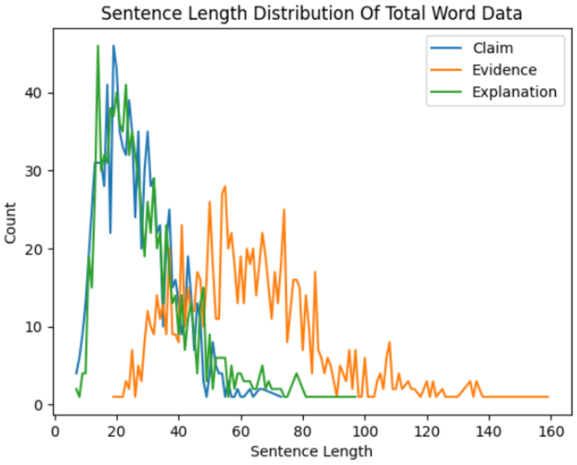
Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901.

A Length Distribution

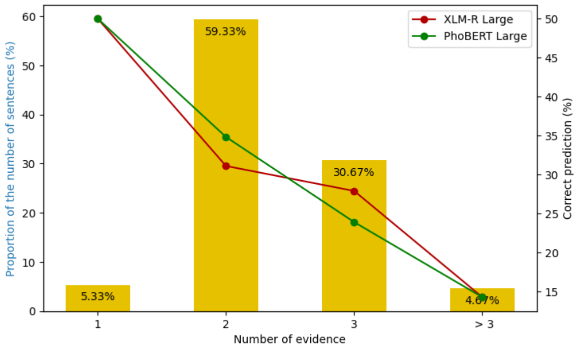
Figure 2 illustrates the length distribution of claims, explanations, and evidence across the entire dataset. We performed word segmentation before length statistics using the VnCoreNLP toolkit (Vu et al., 2018).



Hình 2: Length distribution in ViFC corpus

B Affect of the amount of evidence in prediction veracity of claim

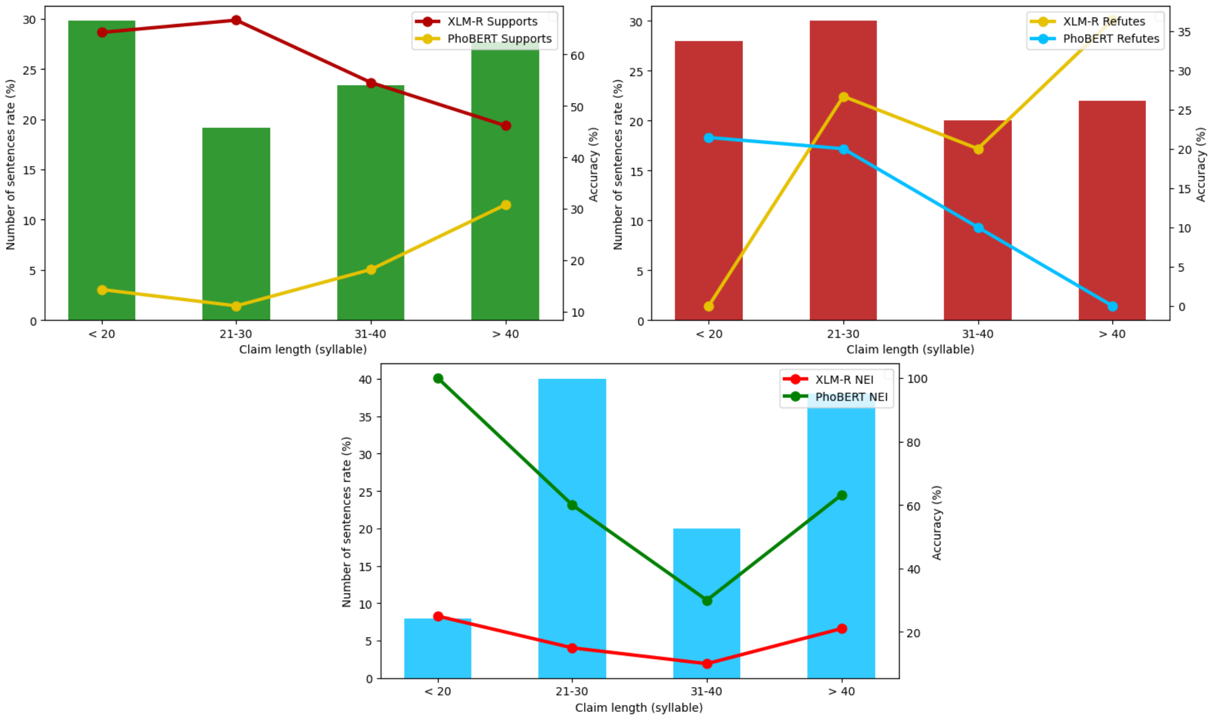
We analyzed the model's prediction results when increasing the number of evidence sentences for the two best-performing models in the verdict prediction task (see section 5.1 for details) on Dev set. The results are presented in Figure 3.



Hình 3: Affect of the amount of evidence in prediction veracity of claim

C Affect of Claim Length

To understand the impact of the claim length on the model's prediction results, we selected the two models that achieved the highest performance in the verdict prediction task (see section 5.1). We analyzed the prediction results for each claim length on the Dev set. The prediction results are illustrated in Figure 4.



Hình 4: Affect of claim length