ViFactCheck: Empowering Vietnamese Fact-Checking across Multiple Domains with a Comprehensive Benchmark Dataset and Methods

Anonymous ACL submission

Abstract

With the rapid development of online information platforms, barriers to the dissemination of information, particularly in media, are diminishing. However, this context has led to various issues, including the proliferation of fake news. Thus, a high-quality datasets and robust solutions for fact-checking, especially for low-resource languages, are essential. This study presents the **ViFactCheck** dataset, the first publicly benchmark **Vi**etnamese **Fact-Check**ing dataset for multiple online news domain. Comprising 7,232 human-annotated statements from reputable Vietnamese online news sources, the dataset covers 12 topics and follows a strict data-constructing process. We also evaluate state-of-the-art monolingual and multilingual pre-trained language models on the ViFactCheck dataset. On the ViFactCheck dataset, the XLM-R_{large} model outperforms robust baseline models such as mBERT, XLM-R_{base}, PhoBERT$_{large}$, PhoBERT$_{base}$, ViBERT achieving a notable macro F1 score of 78.40%. These findings demonstrate the dataset's potential for practical applications.

1 Introduction

The communication landscape has undergone a profound transformation, resulting in the rapid proliferation of communication tools. This transformation has not only revolutionized how we convey information but has also significantly impacted global knowledge consumption. The rapid evolution of contemporary communication methods has led to an immense influx of information, which has become a valuable resource for individuals, businesses, and governments worldwide.

However, this information explosion has given rise to several issues, the most glaring of which is the rampant dissemination of false news (Shu et al., 2017). Furthermore, extensive research Vosoughi et al. (2018) reveals the alarming pace at which false information spreads, often surpassing the reach of legitimate content and presenting significant concerns.

The propagation of disinformation, rumors, and fake news poses a serious threat to societies and public discourse (Olan et al., 2022). The behavior of internet users who share news based solely on headlines, without delving into the substantial content of articles, is a major driver of misinformation. As the volume of data requiring validation across various online platforms and non-mainstream sources continues to grow, the need for information and news verification has become increasingly critical.

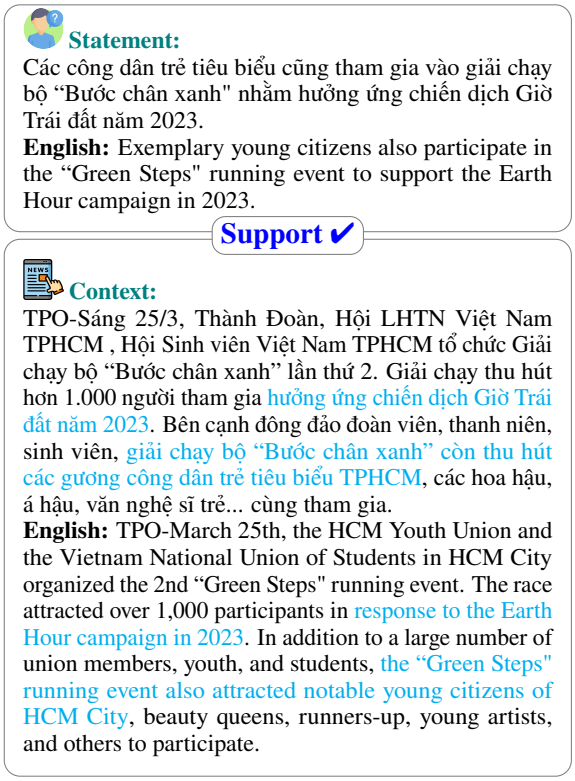


Figure 1: A instance of the Vietnamese fact-checking task. The blue-highlighted words serve as persuasive evidence in determining the label (Support) assigned to the statement. For brevity, only the relevant snippet of the document is shown.

Fact-checking, a rigorous process of verifying the accuracy of statements in specific contexts, relies on informed individuals using evidence, reasoning, and available information to make well-founded judgments (see Figure 1). While substantial efforts have been devoted to statement verification datasets in English (Thorne et al., 2018; Aly et al., 2021; Schuster et al., 2021), resources for fact-checking in low-resource languages like Vietnamese are limited. This scarcity is primarily due to the absence of guidance resources for analyzing the structure and semantics of Vietnamese sentences.

This study introduces the development of ViFactCheck, a pioneering publicly available benchmark fact-checking dataset for Vietnamese online news, covering multiple domains. Our main contributions are described as follows:

- Constructing ViFactCheck, a first human-generated benchmark fact-checking dataset on Vietnamese. ViFactCheck covers 12 popular domains of Vietnamese online news. This dataset consists of 7,232 statements that have undergone rigorous quality control measures, ensuring the highest quality of the dataset.
- Conducting various experiments employing several state-of-the-art language models including XLM-R, ViBERT, PhoBERT, and multilingual BERT on the ViFactCheck dataset. These models have been fine-tuned and evaluated to investigate their effectiveness for the task.
- Undertaking a comprehensive analysis of the limitations and challenges encountered during the development of the ViFactCheck dataset, providing valuable insights to guide future research endeavors.

The structure of this work is organized as follows: Section 2 provides an in-depth overview of literature relevant to the Fact-Checking task. Section 3 presents the comprehensive process of constructing ViFactCheck benchmark dataset. Following that, Section 4 demonstrate the result of various experiments and identify challenges. Finally, Section 5 concludes the study and outlines future research directions.

2 Related Works

2.1 Benchmark Datasets for Fact-Checking

This section investigates the landscape of fact-checking datasets, building upon the research of Hu et al. (2022), and classifies them into two main categories: English and non-English. An overview of these datasets is presented in Table 1.

In the field of fact-checking, several English datasets have attracted significant attention, playing an essential role in advancing research in this domain. Notably, the FEVER dataset (Thorne et al., 2018), VitaminC (Schuster et al., 2021), and LIAR dataset (Wang, 2017) have emerged as notable benchmarks, acclaimed for their comprehensive information and extensive scale. These datasets sourced from reputable platforms like Wikipedia[1], offer a rich resource for fact-checking efforts. Furthermore, certain datasets, such as MultiFC (Augenstein et al., 2019), LIAR (Wang, 2017), and Snopes (Hanselowski et al., 2019) created from various fact-checking websites, contribute to the diversity and authenticity of the data.

Non-English fact-checking datasets, on the other hand, have various limitations due to limited resources as compared to their English equivalents. The DANFEVER (Nørregaard and Derczynski, 2021) and ANT (Khouja, 2020) dataset, for example, were constructed by modifying sentences from Arabic news and Danish Wikipedia, respectively. The Chinese CHEF dataset (Hu et al., 2022) containing a more substantial collection of 10,000 real-world statements, each carefully guided by annotated evidence. In addition, the multilingual xFACT dataset (Gupta and Srikumar, 2021) is a significant resource, providing fact-checking data in 25 languages. It is important to note that the xFACT dataset is primarily concerned with building a multilingual model, which results in comparatively smaller datasets for each specific language.

2.2 Existing Approach for Fact-Checking

The fundamental approach to the fact-checking task involves binary classification, where statements are classified as either true or false (Potthast et al., 2018; Nakashole and Mitchell, 2014). Building upon this fundamental, Schuster et al. (2021) presents a more comprehensive perspective by proposing a multi-class classification approach, wherein statements are classified as Support, Refuted, or Not Enough Information (NEI).

Various methods have been explored to address the fact-checking task, with neural semantic matching network (Nie et al., 2019), graph modelling (Zhong et al., 2020), and the widely popular Trans-

[1] https://www.wikipedia.org/

Table 1: Overview comparison of typical fact-checking datasets. Real-World denotes that the dataset contain statements generated by human and the mentioned event is indeed real in Real-World.

	Dataset	Domain	Labels	# Claims	Real-World	Language	Source	#Evidence
English	FEVER (Thorne et al., 2018)	Multiple	3	185,445	✗	English	Wikipedia	Multi
	FEVEROUS (Aly et al., 2021)	Multiple	3	87,026	✗	English	Wikipedia	Multi
	VitaminC (Schuster et al., 2021)	Multiple	3	488,904	✗	English	Wikipedia	Single
	MultiFC (Augenstein et al., 2019)	Multiple	2-40	36,534	✔	English	Fact-check	Multi
	LIAR (Wang, 2017)	Multiple	6	12,836	✔	English	Fact-check	W/O
Non-English	CHEF (Hu et al., 2022)	Multiple	3	10,000	✔	Chinese	News/Fact-check	Multi
	DANFEVER (Nørregaard and Derczynski, 2021)	Multiple	3	6,407	✗	Danish	Wikipedia	Multi
	ANT (Khouja, 2020)	Multiple	2	4,547	✗	Arabic	News	Multi
	ViFactCheck (Ours)	Multiple	3	7,232	✔	Vietnamese	News	Multi

former based pre-trained language (Vaswani et al., 2017) emerging as robust solution due to their outstanding performance.

Among these approaches, the BERT model (Devlin et al., 2019) has attracted considerable attention. Soleimani et al. (2020) employed BERT to address the fact-checking task, leveraging it on the FEVER dataset (Thorne et al., 2018). Similarly, Liu et al. (2020) utilized the kernel graph attention network in conjunction with BERT models, including BERT$_{base}$, BERT$_{large}$, and RoBERTa$_{large}$. Additionally, Nørregaard and Derczynski (2021) employed a range of multilingual models, such as mBERT, XLM-R$_{base}$, XLM-R$_{large}$, and mBERT for Danish on the DanFEVER dataset.

2.3 Fact-Checking in Vietnamese

To the best of our knowledge, there is no publicly Vietnamese fact-checking dataset available that has been specifically developed and tailored to meet the needs and requirements of the Vietnamese. Previously, Duong et al. (2022) proposed a method that combines knowledge graph (KG) and BERT for fact-checking task on the Vietnamese dataset. This research used a dataset consisting of 129,045 triples extracted from Wikipedia. However, this dataset has not been open for research.

Specifically, the ViNLI dataset (Huynh et al., 2022) focuses on natural language inference tasks for the Vietnamese. It serves as a valuable resource for enhancing language comprehension and understanding in Vietnamese contexts. However, a limitation of this dataset is that the inferred sentences are still rewritten based on one specific phase from the paper. This leads to the fact that the challenges of the ViNLI dataset have not met the requirements of fact-checking task. This absence poses a challenge in ensuring the accuracy and reliability of information in the Vietnamese context.

3 ViFactCheck Dataset's Creation Process

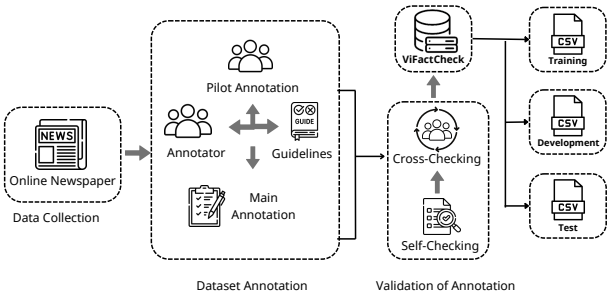


Figure 2: The ViFactCheck dataset contruction process.

Figure 2 presents the process of constructing ViFactCheck, the first fact-checking benchmark dataset in multiple domains of Vietnamese news. Our data construction process consists of three phase: Data collection, Dataset annotation, and Validation of annotation. Each phase was strictly monitored by experts to ensure the high-quality of the dataset.

3.1 Data Collection

The data used for this study was collected from reliable online newspaper websites in Vietnam. These websites are government-licensed, have huge visitor counts, and provide up-to-date news. Data was collected from the following newspapers: Bao Chinh Phu, VnExpress, Dan Tri, Nguoi Lao Dong, Tuoi Tre, Tin Tuc, Phap Luat HCM, Thanh Nien, and Tien Phong. Refer to the Appendix B for further information on these newspapers.

To extract news articles from these online newspaper websites, we utilized two Python libraries, namely BeautifulSoup[2] and Selenium[3]. These libraries are well-known for their robust capabilities

[2]https://pypi.org/project/beautifulsoup4/
[3]https://pypi.org/project/selenium/

in data collection from websites. Following the approach by Kotonya and Toni (2020), we crawled the full text of each news article, which includes the Title, Content, Topic, Description, and URL.

A key consideration in our data collection process was to ensure the dataset remains current and reflective of the present news landscape. Therefore, we specifically gathered articles published between February and March 2023. This meticulous selection approach guarantees that our dataset accurately captures the prevailing state of affairs during that time period. In total, we collected 1,000 articles, covering a diverse range of 12 popular topics.

After we finished collecting data, we discovered that the article descriptions also give useful information. As a result, we combined the content and description sections into one field called "context", which contains the whole context of each article. This revised dataset will serve as the foundation for our research and analysis.

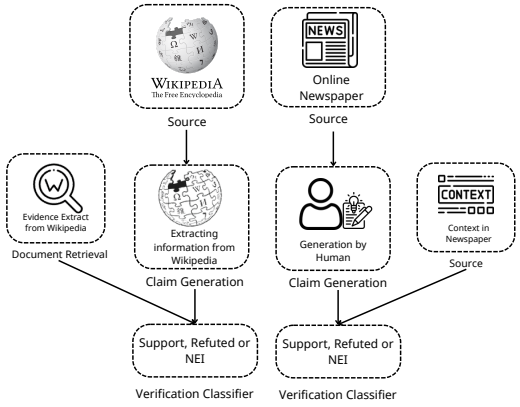
Overall, the combination of reliable news sources, rigorous collecting methodologies, and a targeted up-to-date time period ensures the reliability and relevance of the dataset, making it a valuable resource for research on Vietnamese.

3.2 Dataset Annotation

In the dataset annotation phases, we use the Label Studio[4], an open-source platform that provides an intuitive interface and supports many labeling tasks across various types of data.

To ensure linguistic proficiency and cultural context, we enlisted the expertise of seven university students, native speakers of Vietnamese, who exhibited exceptional command over the language. These proficient annotators received appropriate remuneration at a rate of 0.032 USD per statement, acknowledging the significance of their contributions to the annotation endeavor. Comprehensive guidelines were provided to the annotators to ensure a cohesive and systematic approach:

(1) The annotation process consisted of the generation of six statement pairs for each article in the dataset, resulting in two pairs for each designated label—Support, Refuted, and NEI (Not Enough Information). **(2)** For the Support and Refuted labels, annotations were grounded in the intrinsic information and contextual evidences derived directly from the corresponding news. The NEI label, on the other hand, needed a more nuanced approach, requiring the addition of external information and context, which might either align with the truth or deviate from it. **(3)** The generated statements must adhere to certain rules: attempting to paraphrase the sentences in the article, inferring the statement by combining multiple pieces of information, and meticulously avoiding spelling and abbreviation errors that could harm the dataset's quality. **(4)** To enrich the dataset with diverse perspectives and challenges, annotators were encouraged to leverage their broad vocabulary and skilled sentence-writing techniques, thereby introducing valuable nuances into the annotations.



(a) Thorne et al. (2018). (b) Our proposed process.

Figure 3: Statement Labeling Pipeline in the FEVER dataset and the ViFactCheck dataset.

Unlike prior datasets such as ANT (Khouja, 2020) and DANFEVER (Nørregaard and Derczynski, 2021), which inherited their data constructing process from Thorne et al. (2018), our methodology was innovatively adapted to the domain of Vietnamese online news data. As shown in Figure 3b, the key component of statement generation involves human annotators, who expertly extracted insights from the facts and contextual nuances within the news. Following that, each statement was meticulously assigned its proper label, guided by the contextual information incorporated within the relevant article (see Appendix F).

This methodological enhancement was fundamentally inspired by the awareness that online news data shows specific complexities and nuances, needing an annotation technique that accurately captures and mirrors its nuanced nature. Additionally, human-generated statements are more recognizable and natural than information extraction, allowing them to cover more cases and necessitate higher levels of inference using several pieces

[4] https://labelstud.io/

of evidence. By rigorously aligning the dataset construction with the substance of online news, we ensured its enhanced relevance and efficacy in guiding fact verification and advancing related research efforts.

3.2.1 Pilot Annotation

The first stage of dataset annotation is the pilot annotation, which is used to familiarize annotators with the statement generation and verification classifier process described above.

We conducted a pilot annotation with each annotator placing 120 statements corresponding to 20 randomly selected articles from the dataset. The annotators were instructed to proofread carefully and rigorously adhere to the annotation guidelines that we sent earlier. The annotators were encouraged to use their own vocabulary and diversify their sentence structures. Finally, to check the pilot annotation process, we reviewed the statements and labels of the statements. High expert provided detailed feedback and asked annotators to revise any details or labels that do not met the requirements with the annotation guidelines.

3.2.2 Main Annotation

To ensure an efficient and coherent annotation process, we divided the dataset into seven distinct, non-overlapping subsets. Each annotator, having already gained familiarity with the task during the pilot annotation phase, was assigned one subset for comprehensive annotating. Throughout the annotation process, strict adherence to the guidelines was emphasized to maintain consistency and uphold the dataset's overall quality.

Writing statement: Before writing any statement, annotators meticulously proofread the article, ensuring a comprehensive understanding of its information, which often comprises multiple aspects. By grasping the article's content, annotators expertly combined relevant information to generate reasoning statements in line with the definitions of the three labels-Support, Refuted, and NEI (Not Enough Information). This meticulous adherence to guidelines resulted in accurately and contextually appropriate statements, enhancing the dataset's quality and facilitating valuable contributions to stance classification research.

3.3 Validation of Annotation

After completing the main annotation phases, to ensure the quality and consistency of the dataset, we perform self-checking and cross-checking: **(1)** Self-checking: annotators review their statements and labels, checking for grammar errors and typos. **(2)** Cross-checking: annotators cross-check each other's work. If any mistakes are found in the dataset, they discuss and correct them together. In addition, we follow the success of FEVER dataset (Thorne et al., 2018) and utilized the Fleiss Kappa measure to assess inter-rater agreement.

Metric For Inter-Annotator Agreement: Fleiss Kappa is commonly used to evaluate inter-annotator agreement (IAA) in several tasks and is widely considered as the benchmark (McHugh, 2012). As a result, we employ Fleiss Kappa (Fleiss, 1971) to compute inter-annotator agreements of annotators and quality assurance of human annotation. Fleiss Kappa can be formulated as follows:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\bar{P}$ represents the observed overall agreement, and $\bar{P}_e$ represents the expected mean proportion of agreement due to chance.

We randomly selected 10% of the statements (n = 726) from the labeled dataset and formed a group of three annotators to re-label the statements, which were written by a different individual (annotated labels were concealed). The inter-rater agreement was then calculated using the Fleiss Kappa measure for three classes (Support, Refuted, and NEI). We achieved a consensus level of 0.83, showing a very high level of agreement and indicating that the quality of the dataset is sufficiently high.

3.4 Data Analysis

Dataset basic statistic. The ViFactCheck dataset contains 7,232 samples divided into three subsets: training, development, and test with a ratio of 7:1:2. Basic statistics of the three subsets are shown in Table 2. We found that the average length of a context is around 700 words, with the longest one being 3,602 words. The richness of the context is highly beneficial for models with large parameters, such as XLM-R, as they can capture the maximum features of the data. On average, each sentence in the statement falls around 36 words, with the maximum being 165 words. One notable difference of the ViFactCheck dataset compared to other datasets such as CHEF (Hu et al., 2022) and FEVER (Thorne et al., 2018) is the length of the context. Additionally, another significant difference is that the length

of the statement in ViFactCheck is considerably higher than in the two aforementioned datasets.

Table 2: Basic dataset statistic of ViFactCheck dataset. The vocab size and length are computed at word level.

		Training	Development	Test
Context	Total samples	1035	496	758
	Avg length	693.2	670.2	690.5
	Max length	3602	2534	3602
	Min length	71	71	71
	Total vocab size	25,382	16,522	21,263
Statement	Total samples	5062	723	1447
	Avg length	35.9	35.6	35.8
	Max length	165	145	135
	Min length	7	10	7
	Total vocab size	12,189	4,555	6,711

Words overlap, new word ratio analysis. Based on prior research such as IndoNLI (Mahendra et al., 2021), ViNLI (Huynh et al., 2022), we employed metrics analogous to the Jaccard similarity to analyze word overlap, calculating the order-agnostic word overlap rates of hypothesis pairs, as well as utilizing the Longest Common Subsequence (LCS) to observe the ordered word overlap[5].

Table 3: Words overlap, new word ratio between statements and contexts in ViFactCheck datasets.

	Support	Refuted	NEI
Jaccard Similarity (%)	11.65	11.19	11.11
Longest Common Sequence	20.31	17.75	19.5
New Word Ratio (%)	6.61	11.51	11.89

In the ViFactCheck dataset, we computed the Jaccard coefficient, LCS, and novel word ratios for sentences human-generated based on tokens. The detailed results are presented in Table 3. The Jaccard and LCS lengths are both low and comparable in the ViFactCheck dataset, and they are lower than those in the previous ViNLI dataset (Huynh et al., 2022), indicating that the dataset we have created exhibits interesting characteristics compared to traditional NLI datasets. Furthermore, prior research by McCoy et al. (2019) has also shown that low overlap ratios pose challenges for models and require higher inference capabilities.

[5]Note that vocabulary size and comment length are computed at the word level.

4 Experiment and Results

4.1 Experimental Configures

All the baseline models are trained and finetuned using AdamW optimization function (Loshchilov and Hutter, 2019). We employed a P100-GPU setup equipped with 16GB of memory to fine-tune baseline models on the ViFactCheck dataset, requiring a total of five days to complete all the experiments we conducted. The hyper-parameters of mBERT, ViBERT, PhoBERT, and XLM-R are set up as follow: *learning_rate = 5e-06, dropout = 0.3, batch_size = 8, epochs = 10*. In particular, for the PhoBERT model, the input text data must be word-segmented (Nguyen and Tuan Nguyen, 2020). Therefore, we used the VnCoreNLP toolkit (Vu et al., 2018) to perform word-segmentation as proposed by the authors of the PhoBERT model.

4.2 Experimental Results

Table 4 displays the performance of the models on the ViFactCheck test dataset. We assessed their performance primarily using the $F1_{marco}$ metric, which combines Precision and Recall. The XLM-R_{large} model excelled on the ViFactCheck dataset, achieving 78.40% accuracy on the test set. When examining individual labels, XLM-R_{large} consistently outperformed other models, obtaining the highest $F1_{marco}$ scores in all categories: 84.16% for Support, 73.92% for Refuted, and 77.13% for Not Enough Information (NEI).

Among the monolingual models, $PhoBERT_{large}$ proved to be a competitive choice, achieving a solid $F1_{marco}$ score of 71.56%. This underscores the proficiency of monolingual in handling Vietnamese fact-checking tasks.

An interesting observation is the notable expertise of all models in predicting the Support label. The context-rich nature of this label significantly contributes to improved prediction accuracy.

Single-Evidence: Notably, the XLM-R_{large} model stood out with an $F1_{macro}$ score of 68.01%, demonstrating its superior effectiveness compared to other models. Additionally, the models exhibited stronger performance in identifying Support and NEI claims compared to Refuted cases. The XLM-R_{base} model, however, underperformed, particularly in detecting Refuted claims with a score of 20.77%. These performance differences raise questions about the inherent model architecture and the impact of their training datasets.

Table 4: Experimental results of multilingual versus monolingual models on ViFactCheck dataset.

	Model	$\text{Test}_{Overall}$				$\text{Test}_{Single-Evidence}$				$\text{Test}_{Multi-Evidence}$			
		F1_{macro}	Support	Refuted	NEI	F1_{macro}	Support	Refuted	NEI	F1_{macro}	Support	Refuted	NEI
Multi	mBERT$_{cased}$	66.10	73.50	61.55	63.24	58.40	72.00	36.08	67.13	49.43	63.88	32.37	51.98
	XLM-R$_{base}$	69.20	75.89	64.08	67.62	53.05	71.45	20.77	66.95	46.38	60.32	28.84	49.98
	XLM-R$_{large}$	**78.40**	**84.16**	**72.92**	**77.13**	**68.01**	**78.62**	**57.64**	**67.78**	**66.60**	**72.24**	**64.00**	**63.57**
Mono	PhoBERT$_{base}$	68.86	78.36	62.07	66.15	54.91	71.10	28.54	65.10	49.02	66.22	23.49	57.35
	PhoBERT$_{large}$	71.56	78.76	64.35	71.57	64.09	77.87	41.12	70.30	63.00	74.01	54.86	60.14
	ViBERT$_{cased}$	54.52	66.02	49.83	47.40	52.62	68.89	28.18	60.78	43.08	61.77	27.42	40.08

Multi-Evidence: The XLM-R$_{large}$ model continued to dominate with an F1_{macro} score of 66.60%. Nevertheless, the performance gap between models appears to be narrowing, indicating that multi-evidence scenarios level the playing field to some extent. Notably, while the Support scores remained consistently high across models, the Refuted scores experienced a notable decline, suggesting challenges in refutation detection in multi-evidence contexts. The NEI scores also indicated potential for improvement, with PhoBERT$_{large}$ showing promise with a score of 60.14%. As multi-evidence scenarios closely resemble real-world situations, improving model performance in this category is crucial. This data highlights the need for further research to optimize model architectures and training methods to enhance efficiency in multi-evidence verification tasks.

4.3 Human Performance

To evaluate human performance in the fact-checking process, we engaged three native Vietnamese-speaking students. They were tasked with annotating a representative subset, which consisted of 200 samples. Notably, these participants had no prior exposure to the task of fact-checking. To ensure their comprehension, they received comprehensive instructions, clarifications on label significance, and additional information to aid them in determining appropriate labels for each sample. The final label was determined through a majority consensus among the assessors.

The results in Table 5 reveal that the top-performing model, XLM-R$_{large}$, has not yet achieved parity with human performance, displaying a disparity of approximately 10%. This underscores the potential for enhancing the model's performance and underscores the complexity of the task. Moreover, human performance on the ViFactCheck dataset stands at 84.93%, which is lower than that observed in other Vietnamese inference datasets like ViNLI (Huynh et al., 2022), VIMQA (Le et al., 2022), ViNewsQA (Van Nguyen et al., 2022), and fact-checking datasets including HoVER (Jiang et al., 2020). This underscores the formidable challenge and complexity associated with the ViFactCheck dataset.

Table 5: Evaluation results of human performance compared to the models on the test set of 200 samples.

Model	F1_{macro}	Support	Refuted	NEI
mBERT$_{cased}$	66.94	71.79	61.84	67.18
XLM-R$_{base}$	66.33	71.64	64.97	62.39
XLM-R$_{large}$	74.95	76.47	73.02	75.36
PhoBERT$_{base}$	71.29	75.19	63.89	74.80
PhoBERT$_{large}$	73.08	79.70	62.30	77.24
ViBERT$_{cased}$	55.66	68.70	48.28	50.00
Hiệu suất con người	**84.93**	**81.25**	**80.95**	**82.38**

4.4 Analysis and Discussion

In order to gain comprehensive insights into the performance of the models, we conducted an in-depth analysis based on various factors, including the length of the context, the topic of the news, and the size of the training dataset.

Effects of context-length. We initiated our investigation by analyzing the test results with respect to the context length (see Figure 4). Notably, the XLM-R$_{large}$ model consistently outperformed all other models in terms of performance across various context lengths. Nonetheless, the context length range of 0-100 included a limited amount of data, resulting in very volatile performance across the models. As the context length within the range of 100-400, the performance of most models improved. Interestingly, in the context length range of 400-1500, there was a decline in performance, particularly within the context length range of 400-500. This observation indicates that longer context tend to negatively impact performance, as they typically contain a wealth of information, making

inference more challenging for the models.

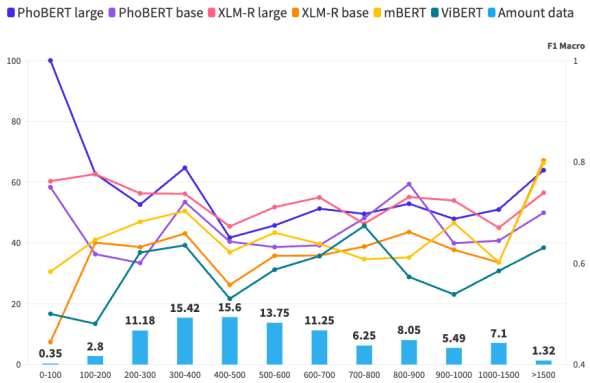


Figure 4: The effect of the length context on test set.

Effects of topic. Next, we investigated the impact of topics on model performance, as shown in Figure 5. Notably, topics such as World and Politics consistently performed well across all models, which is due to the existence of well-structured sentences with fewer factual mistakes or distortions in political contexts. Law, Science, and Culture, on the other hand, performed comparably poorly due to their intrinsic complexity, creating severe challenges for model inference.

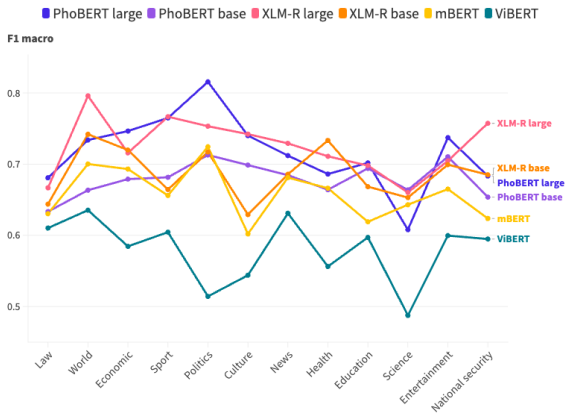


Figure 5: The effect of the topic on the test set.

Effects of the training data size. Finally, to investigate the effect of training data size on model performance, we conducted experiments using various subsets of data, including 1000, 2000, 3000, 4000, and 5062 data points. Figure 6 provides a visual representation of the evaluation performance on these subsets. It is noteworthy that models such as PhoBERT$_{large}$, PhoBERT$_{base}$, XLM-R$_{large}$, and mBERT showcased improved performance as the dataset size increased.

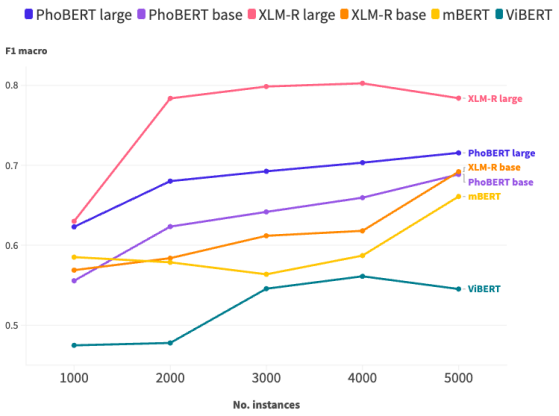


Figure 6: The impact of training data size on test set.

In summary, the comprehensive analysis sheds light on the multifaceted factors influencing model performance. The stability and overall proficiency of XLM-R$_{large}$ across different context lengths underscore its effective for fact-checking tasks. Additionally, the disparities in performance across various topics highlight the challenges caused by complex subjects like Law, Science, and Culture. Moreover, our findings show that increasing the training data size improves the performance of monolingual models like PhoBERT$_{large}$ and PhoBERT$_{base}$, emphasising the need of a robust and diverse training dataset to achieve effective fact-checking results.

5 Conclusion and Future Works

In this study, we introduced ViFactCheck, the first publicly available benchmark for Vietnamese multi-domain fact-checking. With 7,232 samples covering 12 popular topics, ViFactCheck offers a robust dataset to evaluate the performance of various state-of-the-art baseline models. Through a comprehensive analysis, we discovered valuable insights into its limitations and encountered challenges, providing a solid foundation for future research efforts. We truly believe that ViFactCheck will engender new challenges and foster advancements in the field of Vietnamese fact-checking.

Future research directions include exploring large language models for low-resource, developing automated evidence extraction (Wang et al., 2021), building end-to-end fact-checking systems for news (Nadeem et al., 2019), and extending cross-lingual (Gupta and Srikumar, 2021) and cross-domain applications (Augenstein et al., 2019). These initiatives hold promise for advancing fact-checking and preventing misinformation effectively.

Limitations and Ethics consideration

The ViFactCheck dataset and methods present a significant advancement in Vietnamese fact-checking; however, certain limitations must be acknowledged. One notable limitation pertains to potential bias introduced during data labeling by human annotators. These biases, whether conscious or unconscious, may impact the fairness and generalizability of fact-checking models trained on the dataset. Addressing this limitation necessitates the implementation of transparent guidelines and rigorous quality control measures to minimize bias and ensure consistency in the annotations.

During the construction of the ViFactCheck dataset, we prioritized ethical principles to protect individuals' rights and privacy. Informed consent was obtained from data contributors, and data privacy regulations were strictly adhered to. We established clear annotation guidelines and conducted regular quality control checks to minimize potential biases. The dataset was anonymized to safeguard the confidentiality of sources and individuals mentioned in the statements. We commit to using the ViFactCheck dataset solely for research purposes, ensuring its reliability and ethical integrity.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 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, Hanoi, Vietnam. 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.

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

Huong To Duong, Van Hai Ho, and Phuc Do. 2022. Vietnamese fact checking based on the knowledge graph and deep learning. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 530–535.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Khang Le, Hien Nguyen, Tung Le Thanh, and Minh Nguyen. 2022. VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6521–6529, Marseille, France. European Language Resources Association.

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

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. 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.

ML McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. 2019. FAKTA: An automatic end-to-end fact checking system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83, Minneapolis, Minnesota. Association for Computational Linguistics.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland. 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. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 422–428, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2022. Fake news on social media: the impact on society. *Inf Syst Front*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

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.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

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.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

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

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

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.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. 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.

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.

A Task Definition

This paper is motivated by the successful approach introduced by Thorne et al. (2018) and aims to develop an advanced automated system for fact-checking and categorizing human-written statements on Vietnamese online news articles. The primary objective of this system is to accurately assign labels to the given statements, classifying them into one of three categories: Support, Refuted, or Not Enough Information (NEI). These labels are assigned based solely on the information extracted from the corresponding news articles.

Input: The input to the system consists of a Vietnamese news, which serves as the primary source of information, along with a human-authored statement that requires verification against the content of the associated news.

Output: The proposed system is designed to assign labels to the given statements, categorizing them as follows:

1. **Support:** Information is confirmed to be correct according to the content.

2. **Refuted:** Information is determined to be inaccurate compared to the content.

3. **Not Enough Info:** Information that is not sufficiently covered by the corresponding news article. Consequently, such statements cannot be definitively verified or refuted based solely on the content provided within the article.

B Data Collection Source

Table 6: Details of the sources and organizations of the online news sites in the ViFactCheck dataset.

Website	Organization	URL
Bao Chinh Phu	Government of Vietnam	https://baochinhphu.vn
VnExpress	MOST Vietnam	https://vnexpress.net
Dan Tri	MOLISA Vietnam	https://dantri.com.vn
Nguoi Lao Dong	HCM City Committee	https://nld.com.vn
Tuoi Tre	HCM Communist Youth Union	https://tuoitre.vn
Tin Tuc	Vietnam News Agency	https://baotintuc.vn
Phap Luat HCM	HCM City People's Committee	https://plo.vn
Thanh Nien	Vietnam Youth Union	https://thanhnien.vn

C Topic Distribution Analysis

ViFactCheck covers 12 popular topics commonly found in newspapers in Vietnam. Particularly, these are topics that are regularly subjected to misinformation. These topics are compiled in Figure 7. "News" is the most frequently appearing topic because it covers updates on social issues and events in daily life. Other topics such as "World", "Education", and "Economics" also hold significant percentages of 12.4%, 12.9%, and 10.9% respectively. On the other hand, "National security" represents the lowest percentage at 2.0%. This can

be attributed to the low number of articles on this topic in real life. However, due to the absolute need for accuracy in the information provided by this topic, we have collected articles related to it.

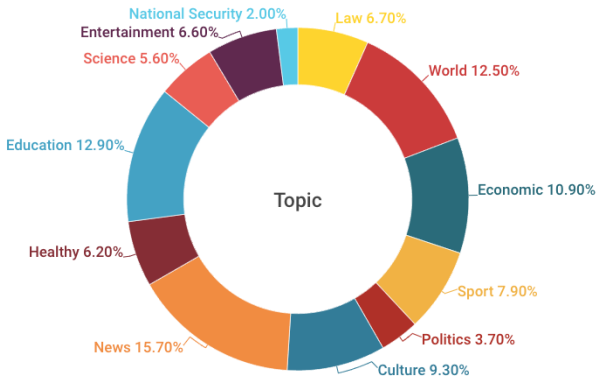


Figure 7: Topic distribution on ViFactCheck dataset.

D Human-Generated Rules

In ViFactCheck datasets, annotators were encouraged to leverage their broad vocabulary and skilled sentence-writing techniques, thereby introducing valuable nuances into the annotations. The basic rules for annotators use of generation are summarized in Table 7.

Table 7: Approaches and rules for generating statements by humans in the ViFactCheck dataset. Denote that a statement can include more than one rules.

	Rules	Ratio (%)
Support	Restructuring the Structure	73.68
	Eliminating or Adding Words	44.21
	Substitute Numbers, Time, or Mathematical Inferences	7.34
	Altering the Word Order in a Sentence	8.42
Refuted	Employing Negation	8.16
	Replacing Words with Antonyms	17.35
	Intentionally misrepresenting quantity	22.45
	Faulty Temporal Logic Structure	16.37
	Erroneous Entity Inference Structure	5.11
	Incorrect Event Inference Structure	47.96
NEI	Infer the sentence with unspecified information.	90.20
	Utilize external knowledge.	10.78

Annotators are required to follow guidelines for creating diverse and challenging data. The distribution of data-generating rule usage for claims related to Support, Refuted, and Not Enough Information (NEI) is shown in Figure 8. To understand how annotators behave in creating ViFactCheck, we analyzed the number of rules used to generate claims. We randomly selected 100 context-claim pairs for Support, Refuted, and NEI.

The primary trend in this dataset reveals an obvious bias for using 1-2 rules, reflecting a standardized annotation process. However, some annotators deviated from this trend, opting for four or more rules, demonstrating an awareness of the data's complexity and diversity. This underscores the importance of judiciously combining rules for reliable and accurate annotation.

The use of multiple rules presents challenges for language model development, introducing complexity into inference and decision-making processes dependent on rule combinations. Nevertheless, it also offers an opportunity to enhance more adaptable language models, ensuring increased accuracy in making inferences.

E Baseline Models

The emergence of transformer-based models, notably BERT variants, has significantly bolstered their efficacy in fact-checking field, demonstrating impressive performance across various datasets (Thorne et al., 2018; Hu et al., 2022; Nørregaard and Derczynski, 2021). As a result, we decided investigating BERT variants to evaluate their effectiveness in the Vietnamese fact-checking task.

Multilingual BERT (mBERT) (Devlin et al., 2019) is a transformer-based model trained on an extensive corpus of 104 languages, including Vietnamese. Its linguistic versatility empowers mBERT to comprehend multiple languages, making it invaluable for fact-checking tasks involving diverse information sources. Addressing multilingual, mBERT enables comprehensive analysis and serves as an excellent tool for ensuring the credibility of data within the Vietnamese fact-checking.

PhoBERT (Nguyen and Tuan Nguyen, 2020) is a RoBERTa-based model specifically developed for the Vietnamese language. Leveraging the powerful Transformer architecture of RoBERTa (Liu et al., 2019), PhoBERT exhibits a profound understanding of the nuances and context of the Vietnamese language. This linguistic precision proves highly beneficial for the Vietnamese fact-checking dataset, as it can discern subtle language nuances that general models might overlook. With its focus on Vietnamese, PhoBERT delivers exceptional efficiency and accuracy when applied to a corpus of the same language, facilitating high-quality fact-checking within the Vietnamese context.

Cross-lingual Language Model - RoBERTa (XLM-R) (Conneau et al., 2020) is a transformer-

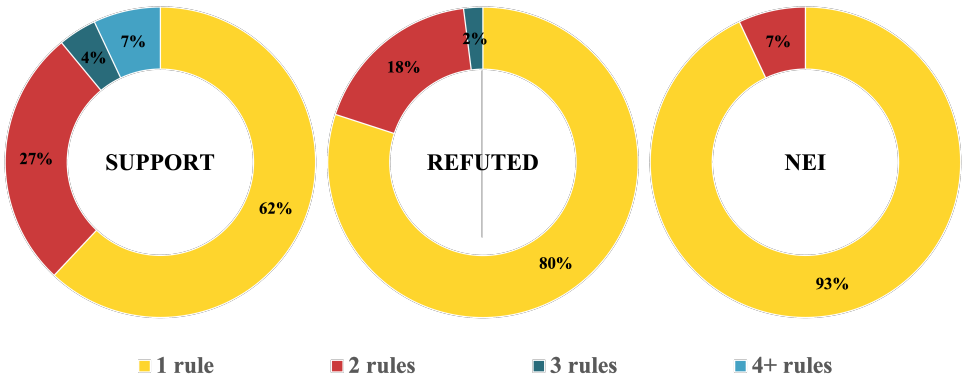


Figure 8: The ratio of combining different rules to create claims in ViFactCheck.

based model trained on 100 languages. This vast linguistic scope means that XLM-R can understand and compare information across different languages. For the fact-checking, this cross-lingual capability proves advantageous, offering a broader context beyond the Vietnamese language. XLM-R's ability to understand and fact-check information from multilingual sources or across language barriers is particularly valuable when dealing with transcends linguistic boundaries.

ViBERT is an architecture based on BERT specifically designed for Vietnamese, was introduced by Bui et al. (2020). Similar to mBERT, ViBERT is pre-trained on a substantial corpus of 10GB of uncompressed Vietnamese text. However, a notable distinction exists between ViBERT and mBERT, ViBERT deliberately excludes insufficient vocabulary, focusing solely on Vietnamese to achieve optimal performance within this language.

By investigating the effectiveness of these BERT variants in Vietnamese fact-checking, we intend to improve the field's abilities in combatting disinformation. The diversity of these models in terms of monolingual understanding, linguistic precision, and cross-lingual capabilities promises to make a significant contribution to the fact-checking landscape, advancing a more credible and precise information ecosystem.

F Data Examples

The ViFactCheck dataset includes various examples of written statements, as illustrated in Table 8. To create a challenging context, annotators were tasked with generating statements based on multiple pieces of evidence (the highlighted words) provided in the context. This approach contributes to the dataset's reliability and enhances its value for fact-checking tasks in the Vietnamese language.

Table 8: Typical samples from the **ViFactCheck** dataset with three labels **Support**, **Refuted**, and **NEI**. The highlighted words is the evidence of the statement.

Context	TPO - Tổng Công ty Cảng Hàng không Việt Nam (ACV) vừa chính thức gia hạn thời gian mời thầu thêm 1 tháng, kéo dài thời gian thực hiện gói thầu thi công nhà ga sân bay Long Thành từ 33 tháng lên 39 tháng. Như vậy, "siêu sân bay" Long Thành sẽ chỉ có thể đưa vào khai thác từ năm 2026 thay vì mục tiêu năm 2025 như trước đó. Tin từ ACV cho hay, đơn vị chính thức điều chỉnh kế hoạch và hồ sơ mời thầu gói thầu thi công xây dựng và lắp đặt thiết bị nhà ga hành khách sân bay Long Thành giai đoạn 1 (do ACV làm chủ đầu tư). Cụ thể, thời gian mời thầu được gia hạn thêm 1 tháng, kéo dài tới sáng ngày 28/4, thay vì tới ngày 28/3 như trước đó. ... Gói thầu thi công nhà ga hành khách sân bay Long Thành trị giá hơn 35 nghìn tỷ đồng do ACV làm chủ đầu tư. Đây là gói thầu lớn nhất dự án sân bay Long Thành... (**English:** TPO - Vietnam Airport Corporation (ACV) has officially extended the bidding period by an additional month, prolonging the implementation time for the construction contract of Long Thanh Airport's passenger terminal from 33 months to 39 months. Consequently, the "mega airport" Long Thanh will only be operational by 2026 instead of the previous target of 2025. According to ACV, the organization has formally adjusted the plan and tender documents for the construction and installation of the passenger terminal at Long Thanh Airport Phase 1 (with ACV as the main investor). Specifically, the bidding period has been extended by one month, now ending on the morning of April 28, instead of the previous deadline of March 28. ... The construction contract for Long Thanh Airport's passenger terminal, valued at over 35 trillion VND is being overseen by ACV. This is the largest contract within the Long Thanh Airport project.)
Support	Việc nhà thầu thi công xây dựng và lắp đặt thiết bị nhà ga hành khách sân bay Long Thành giai đoạn 1 bị điều chỉnh, thời gian bị kéo dài tới sáng ngày 28/4 thay vì tới ngày 28/3 như dự kiến. **English:** *The construction and installation contract for the Long Thanh Airport Phase 1 passenger terminal has been adjusted, with the timeline extended to the morning of April 28 instead of the originally anticipated March 28.*
Refuted	Tổng Công ty Cảng Hàng không Việt Nam (ACV) vừa gia hạn thời gian mời thầu thêm thời gian 2 tháng, tức "siêu sân bay" Long Thành sẽ chỉ có thể đưa vào sử dụng từ năm 2026 thay vì năm 2025 như dự kiến ban đầu. **English:** *Vietnam Airport Corporation (ACV) has recently extended the bidding period by an additional 2 months, meaning that the "mega airport" Long Thanh will only be operational by 2026 instead of the originally planned year 2025.*
NEI	Gói thầu lớn nhất dự án sân bay Long Thành là gói thầu thi công nhà ga hành khách với trị giá hơn 35 nghìn tỷ đồng, được tài trợ bởi công ty Hàn Quốc. **English:** *The largest contract within the Long Thanh Airport project is the construction of the passenger terminal, valued at over 35 trillion VND, and it is sponsored by a South Korean company.*

Table 8 Continued: Typical samples from the **ViFactCheck** dataset with three labels **Support**, **Refuted**, and **NEI**. The highlighted words is the evidence of the statement.

Context	(Dân trí) - Mỗi tháng, Ukraine ước tính dành hơn 3 tỷ USD cho chi tiêu quân sự để đối phó với chiến dịch quân sự đặc biệt của Nga. Binh sĩ Ukraine khai hỏa lựu pháo M777 (Ảnh: Reuters). Tại cuộc họp với Hiệp hội Kinh doanh châu Âu (EBA) hôm 29/3, Bộ trưởng Tài chính Ukraine Sergey Marchenko cho biết, nước này đang chi 130 tỷ hryvnia (3,5 tỷ USD) mỗi tháng cho quân sự. Ngoài ra, theo ông Marchenko, ngân sách của Ukraine nhận khoảng 80 tỷ hryvnia (khoảng 2,2 tỷ USD) mỗi tháng. "Nhiệm vụ chính là tạo điều kiện tài trợ cho quân đội" ông Marchenko nói. ... Chính phủ Ukraine có kế hoạch bù đắp thâm hụt bằng viện trợ từ phương Tây. Mỹ và các đồng minh liên tục hỗ trợ cả về tài chính, nhân đạo và quân sự cho Kiev kể từ khi xung đột ở Ukraine cách đây hơn một năm... (**English:** Each month, Ukraine estimates allocating over 3 billion USD for military expenses to counter Russia's special military campaign. Ukrainian soldiers fire M777 howitzers (Photo: Reuters). During a meeting with the European Business Association (EBA) on March 29, Ukrainian Finance Minister Sergey Marchenko revealed that the country is spending 130 billion hryvnia (3.5 billion USD) monthly on military expenditures. Additionally, according to Marchenko, Ukraine's budget receives around 80 billion hryvnia (approximately 2.2 billion USD) each month. "The primary task is to provide funding for the military," Marchenko said. ... The Ukrainian government plans to offset the deficit with assistance from the West. The United States and its allies have been providing continuous financial, humanitarian, and military support to Kiev since the conflict in Ukraine began over a year ago...)
Support	Nhiệm vụ chính là tạo điều kiện tài trợ cho quân đội nên ngân sách của Ukraine nhận khoảng 80 tỷ hryvnia (khoảng 2,2 tỷ USD) mỗi tháng. **English:** *The primary objective is to facilitate financing for the military, and as a result, Ukraine's budget receives approximately 80 billion hryvnia (around 2.2 billion USD) each month.*
Refuted	Ukraine ước tính mỗi tháng chi 130 tỷ hryvnia, khoảng hơn 4 tỷ USD cho chi tiêu quân sự để đối phó với chiến dịch quân sự đặc biệt của Nga. **English:** *Ukraine estimates spending approximately 130 billion hryvnia, which is over 4 billion USD, on military expenses each month to counter Russia's special military campaign.*
NEI	Nhờ viện trợ từ phương Tây, chính phủ Ukraine có kế hoạch bù đắp thâm hụt, Mỹ và các đồng minh như Nhật Bản, Hàn Quốc,... liên tục hỗ trợ cả về tài chính, nhân đạo và quân sự cho Kiev kể từ khi có xung đột ở Ukraine hơn một năm về trước. **English:** *With assistance from the West, the Ukrainian government has a plan to offset the deficit. The United States and its allies such as Japan, South Korea, and others have been providing consistent financial, humanitarian, and military support to Kiev since the conflict in Ukraine began over a year ago.*