A Hybrid Approach to Building a Vietnamese Fact-Checking Dataset and Benchmarking Language Models

Nhat-Phuong Vo[*], Gia-Minh Hoang[†], Lan-Anh Nguyen Thi[‡], Tien-Dat Nguyen The[§], Kiet-Van Nguyen[¶]
University of Information Technology, VNU-HCM, Vietnam
[*]22521172@gm.uit.edu.vn, [†]21520861@gm.uit.edu.vn, [‡]22520061@gm.uit.edu.vn,
[§]22520225@gm.uit.edu.vn, [¶]kietnv@uit.edu.vn

Abstract

In the digital information age, verifying the accuracy of information has become more critical than ever. This study introduces a novel hybrid approach to constructing training data for Vietnamese fact-checking by integrating synthetic data generated by large language models (LLMs) such as GPT-4 and Gemini Pro with real-world data collected from the social media platform Facebook. The research team developed a dataset comprising 12,000 diverse claim–evidence pairs, including 10,000 systematically generated samples by LLMs and 2,000 carefully curated samples from social media sources to enhance real-world relevance. This dataset was subsequently used to evaluate the performance of five recent state-of-the-art Vietnamese language models: MDeBERTa, PhoBERT, ViBERT, ViT5, and DistilUSE + LR. Results show that MDeBERTa achieved the highest performance, with an F1-score of 94.6% on the test set.

1 Introduction

The rapid development of social media platforms has fundamentally transformed the way people access, share, and engage in discussions about current events, scientific discoveries, technological innovations, and various other domains. On platforms such as Facebook, TikTok, and Twitter, information regarding new scientific findings, astronomical phenomena, technological achievements, or social issues is frequently disseminated at an extraordinary pace, capturing the attention and sparking debates among a wide online audience. However, alongside the benefits of greater information accessibility, the digital environment also poses significant challenges in verifying the accuracy of disseminated content, as misinformation, disinformation, and unverified claims can easily spread and negatively impact public perception.

To address this issue, this study focuses on constructing a Vietnamese fact-checking dataset that spans multiple domains. The claim–evidence pairs are designed around real-world, timely events that are commonly encountered on social media. Based on this dataset, the research develops and evaluates automated fact-checking models, with the overarching goal of contributing to a more transparent, accurate, and trustworthy social media environment for Vietnamese-speaking users.

2 Related Works

Our approach is inspired by the notable contributions presented in the paper *ViFactCheck: A New Benchmark Dataset and Methods for Multi-domain News Fact-Checking in Vietnamese* by Tran Thai Hoa and colleagues (Hoa et al., 2025). While primarily focusing on news articles, ViFactCheck laid the groundwork for building Vietnamese fact-checking systems by providing a standardized dataset consisting of claims, evidence, and classification labels (SUPPORT, REFUTE, NEI). This resource serves as an essential foundation for expanding research into multi-domain datasets, where information is conveyed in diverse forms and through heterogeneous linguistic styles.

Over the past five years, various language models have been applied to the Vietnamese fact-checking task. Among them, several models stand out:

- **MDeBERTa (He et al., 2023)**: A multilingual variant of DeBERTa, pre-trained on the large-scale CC100 corpus. It demonstrates strong performance on multilingual natural language understanding tasks, particularly in zero-shot cross-lingual settings.
- **PhoBERT (Nguyen and Nguyen, 2020)**: The first BERT-based model pre-trained specifically for Vietnamese. It has shown high effectiveness in capturing semantic representations.

- **DistilUSE + LR (Reimers and Gurevych, 2020)**: A hybrid approach combining DistilUSE, a lightweight and efficient multilingual sentence embedding model, with a simple yet robust Logistic Regression classifier. This method facilitates fast and accurate multilingual text classification.

- **ViBERT (Bui et al., 2020)**: A BERT-based model tailored for Vietnamese, achieving strong results across various text classification tasks.

- **ViT5 (Phan et al., 2022)**: A Vietnamese-adapted variant of the T5 model, supporting generative methods for fact-checking applications.

Although most prior studies have focused on mainstream news media data, emerging approaches have started to explore alternative sources beyond journalism—particularly social media platforms, where language is often informal, unstructured, and highly diverse.

To address these challenges, our team constructed a high-practicality, multi-domain Vietnamese fact-checking dataset. The dataset was developed with the assistance of large language models (LLMs), which played a key role in several stages: generating candidate claims, retrieving or generating relevant evidence, and assigning initial labels through a semi-automated process. In addition to fully synthetic data generated by LLMs, we collected a portion of real claims from social media platforms such as Facebook to enhance the dataset's authenticity and diversity. All data samples were subsequently manually reviewed and refined to ensure high consistency and correctness.

Overall, prior research has served as a crucial foundation for our methodology. By building upon state-of-the-art language models and leveraging the generative capabilities of LLMs, we were able to extend the fact-checking task into a more realistic and diverse data space—covering a broad range of topics and contexts in Vietnamese social discourse.

3 Dataset

3.1 Description

The dataset constructed in this study comprises two primary sources: (1) fully synthetic data generated by large language models (LLMs) such as GPT-4 and Gemini 2.5 Pro, and (2) real-world data collected from Facebook, a widely used social media platform in Vietnam.

For the first source, the research team used LLMs to systematically generate claim–evidence pairs along with their corresponding labels, ensuring topic diversity, logical coherence, and meaningful interaction between claims and evidence. More than 10,000 such pairs were created to simulate common fact-checking scenarios encountered in digital media environments.

For the second source, approximately 2,000 real claims were collected from Facebook. GPT-4 was then employed to generate appropriate evidence and assign labels (SUPPORTS or REFUTES). Importantly, these claim–evidence–label triplets were manually verified by human annotators to ensure accuracy and reliability. The data construction process was rigorously designed with strict requirements regarding quality, length, and real-world applicability.

The resulting dataset—named **ViFAiC**—serves as a novel and valuable resource for advancing research and development of Vietnamese fact-checking systems, particularly within the context of modern social media ecosystems.

3.2 Building Dataset

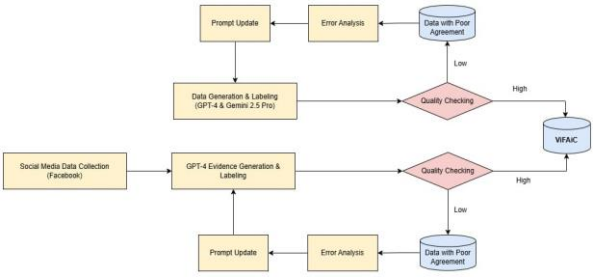

Figure 1: Overview of the ViFAiC dataset construction pipeline.

1. **Social Media Data Collection:** A portion of claims was collected from social media platforms such as Facebook, reflecting trending topics and prevalent public opinions within the community.

2. **Data Generation & Labeling:** Data samples were automatically generated using large language models such as GPT-4 and Gemini 2.5 Pro. In this phase, the system produced claims along with initial classification labels based on pre-designed prompts to ensure content

consistency and alignment. To enhance diversity, a variety of templates and topical domains, such as politics, science, lifestyle, and economics, were used interchangeably in the prompts.

3. **GPT-4 Evidence Generation & Labeling:** For claims collected from social media, corresponding evidence and labels were also generated using GPT-4, with modifications tailored to reflect real-world contexts and informal, everyday language.

4. **Quality Checking:** All data samples underwent a quality review process to assess their reliability and suitability for inclusion in the official dataset. This evaluation considered aspects such as the clarity of the claim, the accuracy, and the relevance of the supporting evidence.

    4.1. **ViFAiC:** If a sample exhibited high agreement, it was included in *ViFAiC* - the official Vietnamese fact-checking dataset.

    4.2. **Data with Poor Agreement:** Samples with inconsistent labeling or insufficient quality were stored in a separate *Data with Poor Agreement* repository for further review and potential refinement.

5. **Error Analysis:** Samples with low agreement were analyzed to identify the sources of labeling errors. These included unclear annotation guidelines, suboptimal prompts, or misunderstandings by the language model. The team also assessed cases where GPT-4 generated irrelevant evidence, demonstrated poor logical consistency, or produced unstable labels across different runs (prompt instability). Common error patterns were documented to guide future improvements to the data generation pipeline.

6. **Prompt Update:** Based on the error analysis, prompts used for data generation were revised and improved to minimize errors in subsequent iterations. For instance, if the model frequently confused factual claims with speculative ones, additional content constraints and specific examples were incorporated into the prompts.

3.3 Example

Figure 2: An example of a data sample labeled as SUPPORTS in the ViFAiC Vietnamese fact-checking dataset.

Figure 3: An example of a data sample labeled as REFUTES in the ViFAiC Vietnamese fact-checking dataset.

4 Methodology

4.1 Proposed System Architecture

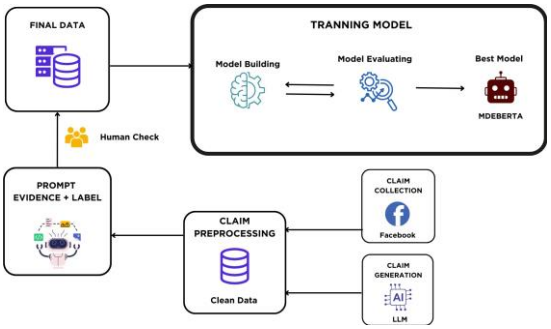

Figure 4: Proposed System Architecture

Description: The proposed system architecture consists of the following components: Data Collection, Data Preprocessing, Label Generation using

LLMs, Manual Verification, Final Dataset Construction, and Model Training. The detailed procedure is as follows:

- **Data Collection:** Data was gathered from two main sources:
  - **Facebook:** A total of 2,000 user comments on various topics were collected from the social media platform Facebook using Selenium.
  - **LLM:** 10,000 claims were synthetically generated using a large language model (LLM).
- **Data Preprocessing:** Claims collected from Facebook were preprocessed to remove noise specific to social media content, such as emojis, teencode, and informal expressions. This step aimed to normalize the text and ensure consistency before generating evidence and labels with the LLM.
- **Label and Evidence Generation using LLM:** Labels and supporting evidence were generated automatically using a large language model (LLM). Each claim was assigned a label based on the relationship between the claim and the evidence: SUPPORTS if the evidence confirms the claim, and REFUTES if the evidence contradicts the claim.
- **Manual Verification:** The automatically generated labels were reviewed and verified by human annotators to ensure accuracy and to prevent errors or biases.
- **Final Dataset Construction:** The verified data served as the final input for model training.
- **Model Training:** A binary classification model was trained, where the input consists of claim–evidence pairs and the output is the assigned label. Among the evaluated models, MDeBERTa achieved the best performance (He et al., 2023).

4.2 Data Processing

4.2.1 Processing Social Media Comments

Data Source: A total of 2,000 comments were collected from posts on the social media platform Facebook.

Objective: To create a Vietnamese fact-checking dataset, where each data sample consists of three components: claim, evidence và label.

Processing pipeline::

1. **Preprocessing and Normalizing Comments:**
   - Remove emojis, emoticons, links, and username tags that include '@[name]'.
   - Replace characters repeated three or more times with only two repetitions.
   - Handle abbreviations, teencode (Vietnamese teen slang), and common English words.
   - Remove redundant whitespace and normalize the output text into a clear and coherent claim.
2. **Generating Evidence and Labeling:** Use the GPT-4 model (via OpenAI API) to generate appropriate evidence for each claim. The model also provides a corresponding label, either SUPPORTS or REFUTES.
3. **Output Requirements:** The evidence must have a clear and direct connection to the claim. It should be 20–100 words long, written in Vietnamese, concise, neutral, and factually accurate. It is encouraged to include specific details such as numbers, locations, dates, or real-world events.

Illustrative Example:

Original comment (Vietnamese):

"T còn đang thực tập cntt không lương đây, mơ 30 củ cho đỡ buồn 😳, cntt mới ra trường 7-10 củ là mừng rớt nc mắt rùi."

Normalized claim (Vietnamese):

```
{
    "claim": "tôi còn đang thực tập Công nghệ thông tin không lương đây, mơ 30 triệu cho đỡ buồn, Công nghệ thông tin mới ra trường 7-10 triệu là mừng rớt nước mắt rồi"
}
```

English translation of claim:

"I'm still doing an unpaid IT internship, dreaming of 30 million just to feel better. New IT graduates

earning 7–10 million is enough to bring tears of joy."

Model Output (Vietnamese):


```
{
    "claim": "tôi còn đang thực tập Công nghệ thông tin không lương đây, mơ 30 triệu cho đỡ buồn, Công nghệ thông tin mới ra trường 7-10 triệu là mừng rớt nước mắt rồi",
    "evidence": "Khảo sát mức lương đầu vào trong ngành CNTT tại Việt Nam cho thấy sinh viên mới tốt nghiệp thường nhận lương khởi điểm từ 7 đến 12 triệu đồng, tùy theo năng lực và công ty. Nhiều sinh viên chấp nhận thực tập không lương để tích lũy kinh nghiệm.",
    "label": "SUPPORTS"
}
```


English translation of evidence:

"Surveys on entry-level salaries in Vietnam's IT sector show that recent graduates typically earn a starting salary of 7 to 12 million VND, depending on their skills and the company. Many students accept unpaid internships to gain experience."

4.2.2 Processing LLM-Generated Data

Data Source: More than 10,000 data samples were generated using large language models such as GPT-4 and Gemini 2.5 Pro.

Objective: To process and normalize the data generated by large language models, including the claim, evidence, and label fields produced by the models.

Processing Pipeline:

1. **Validity Check:** Remove data samples with missing fields (null or empty). Check JSON format and the length of the evidence field. Discard samples with evidence that is too short or too long (outside the 20–100 word range).

2. **Field Normalization:** Normalize the label field into a consistent format: SUPPORTS, REFUTES. Use string operations such as .strip(), .lower(), and .capitalize() to standardize the text.

4.3 Training Models

In this project, we implement and train multiple machine learning and deep learning models to evaluate their effectiveness on the binary classification task ("SUPPORTS" or "REFUTES") based on the (claim, evidence) pair. The models are categorized into two main groups: Machine Learning and Deep Learning.

1) Machine Learning

- **DistilUSE + Logistic Regression (Reimers and Gurevych, 2020):** The *distiluse-base-multilingual-cased-v2* model is used to extract embeddings for each (claim, evidence) pair. This model is kept frozen (i.e., not fine-tuned). The resulting embedding vectors are then passed into a Logistic Regression model for training and classification. This approach is simple yet effective, leveraging the strong semantic representation capabilities of a pre-trained model combined with a linear model that is easy to interpret.

2) Deep Learning

- **MDeBERTa (He et al., 2023):** The mdeberta-v3-base model, a multilingual variant of DeBERTa, is applied for classifying sentence pairs. MDeBERTa combines attention disentanglement and relative position encoding, which enhances deep semantic representation. The model is further fine-tuned on the Vietnamese dataset to adapt to the classification task.

- **PhoBERT (Nguyen and Nguyen, 2020):** PhoBERT is a pre-trained language model for Vietnamese, based on the RoBERTa architecture. In this study, the vinai/phobert-base version is used and fine-tuned for binary classification. Dense layers are added after the output layer, combined with a softmax function to compute classification probabilities.

- **ViBERT (Bui et al., 2020):** ViBERT is a cased BERT model developed by FPT.AI for the Vietnamese language. The FPTAI/vibert-base-cased version is employed for binary classification. The model's output consists of two logits corresponding to the two labels: "SUPPORTS" and "REFUTES".

- **ViT5 (Phan et al., 2022):** ViT5 is a sequence-to-sequence model based on the T5 architecture, developed by VietAI for Vietnamese. The `VietAI/vit5-base` version is fine-tuned to generate output labels ("supports" or "refutes") from input sequences in the format `claim: ... evidence:` Fine-tuning is conducted using the `Seq2SeqTrainer` from the HuggingFace library.

4.4 Model Evaluation

The performance of the models is measured on the *test* set using the following key evaluation metrics:

- **Accuracy**: The proportion of correct predictions.
- **F1-score (macro)**: The harmonic mean of precision and recall, averaged equally across all classes.
- **Precision**: The proportion of correctly predicted positive instances among all instances predicted as positive.
- **Recall**: The proportion of actual positive instances that are correctly predicted.

All models are evaluated under the same conditions to ensure fairness in comparison.

5 Experiments

5.1 Experimental Setup

The experiments are conducted to evaluate the effectiveness of language models in the binary classification task with two labels: SUPPORTS and REFUTES. The dataset is divided into three subsets: training, validation, and test. The training set combines both synthetic data generated by large language models (LLMs) and real-world data collected from social media. The validation set is used to tune model hyperparameters, while the test set is reserved for final evaluation after training.

The models evaluated in the experiments include MDeBERTa, PhoBERT, ViBERT, ViT5, and DistilUSE. Among these, PhoBERT, MDeBERTa, and ViBERT are fine-tuned directly for the classification task. For DistilUSE, the model is used to extract embeddings, which are then fed into a simple classifier such as Logistic Regression. ViT5, on the other hand, is a sequence-to-sequence generation model, where the input is a sentence pair (claim + evidence) and the output is a textual label.

The training parameters are kept consistent across models: the number of epochs is set to 3, batch size is 8, and the learning rate is 2×10^{-5}. The AdamW optimizer is used. Evaluation is performed after each epoch, with macro F1-score selected as the main metric for comparing model performance.

5.2 Experimental Results

After training the models, the team proceeded to evaluate them on the test set. The evaluation metrics included Accuracy, Precision (macro), Recall (macro), and F1-score (macro). The results are presented in Table 1.

Table 1: Evaluation results of the models on the test set

Model	Accuracy	Precision	Recall	F1-score
PhoBERT	0.912	0.911	0.910	0.910
MDeBERTa	0.946	0.945	0.948	0.946
ViBERT	0.839	0.837	0.839	0.838
ViT5	0.923	0.921	0.924	0.922
DistilUSE + LR	0.759	0.760	0.750	0.753

5.3 Analysis and Comparison

From the results table, it is evident that the **MDeBERTa** model (He et al., 2023) achieves the most outstanding performance, with an **F1-score of 94.6%**, demonstrating superior capability in handling Vietnamese text classification tasks. This achievement is largely attributed to the advanced DeBERTa architecture, which is well-known for its strong semantic representation and deep contextual understanding, particularly effective for the linguistic characteristics of Vietnamese.

The **PhoBERT** model (Nguyen and Nguyen, 2020) also attains impressive results, achieving an **F1-score of 91.0%**, reflecting the advantage of being pre-trained specifically on Vietnamese language data. Meanwhile, **ViT5** (Phan et al., 2022), a sequence-to-sequence model not originally designed for classification tasks, still delivers high performance (**F1-score: 92.2%**), indicating the flexible adaptability of the T5 architecture across diverse NLP problems, including text classification.

In contrast, **ViBERT** (Bui et al., 2020) records a lower performance, with an **F1-score of only 83.8%**. This may stem from limitations in the pre-training process or an insufficient scale of training data to fully optimize performance for Vietnamese. Finally, the **DistilUSE + LR** model (Reimers and Gurevych, 2020), although simpler, still produces stable results (**F1-score: 75.3%**), making it suitable

for applications with constrained computational resources.

In summary, **MDeBERTa** stands out as the most effective model among the architectures tested, followed by **ViT5** and **PhoBERT** respectively.

6 Result Analysis

6.1 Analyzed Errors

Based on the results obtained from the best-performing model, **MDeBERTa** (He et al., 2023), the team conducted a detailed analysis and constructed the confusion matrix on the test set, as shown in Figure 5.

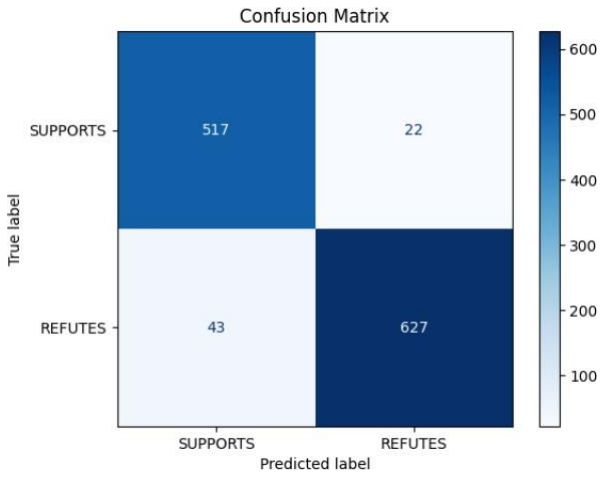

Figure 5: Confusion Matrix

The confusion matrix illustrates that the **MDeBERTa** model achieves relatively strong classification performance, with high accuracy in predicting both labels. Specifically, the model correctly predicts 517 instances of the SUPPORTS label and 627 instances of the REFUTES label. However, some misclassifications remain, predominantly within the REFUTES label, with a total of 43 cases mistakenly predicted as SUPPORTS, compared to 22 SUPPORTS cases misclassified as REFUTES. This indicates that the model tends to be *more prone to confusion when distinguishing claims refuted by evidence*, particularly in instances involving ambiguous semantics or implicit information.

The analysis process revealed two main categories of noteworthy errors:

1. **Errors due to misunderstanding implicit semantics or real-world knowledge**
   In some cases, the model makes incorrect predictions because it lacks real-world knowledge or fails to fully understand the relationship between the *claim* and the *evidence*, even though both are clear to humans. The following example illustrates such a case:

   {
   "claim": "Sai nha, bò tót bị mù màu đỏ và xanh lá cây"
   "evidence": "Bò tót không phân biệt được màu đỏ và xanh lá vì chúng có khiếm khuyết về thị giác màu sắc, nên không bị kích thích bởi màu sắc này."
   "true-label": "SUPPORTS"
   "predicted-label": "REFUTES"
   }

   This claim aims to refute a misconception, but the model fails to recognize that the *evidence* fully supports this refutation. The error may be due to the *claim* starting with a negation phrase ("Sai nha"), which confuses the model into judging the two sentences as contradictory, while in fact, they both affirm that bantengs cannot distinguish red and green colors.

2. **Errors due to misunderstanding the logical relationship between claim and evidence**
   Another common error is that the model assigns the label SUPPORTS when the *evidence* clearly refutes the *claim*, simply because of some shared keywords or vague associations. This often happens when the *evidence* contains soft content or leans towards social or cultural analysis. For example:

   {
   "claim": "Chơi game nhiều là hư hỏng, vô bổ"
   "evidence": "Chơi game quá mức có thể gây hại, nhưng chơi game điều độ có thể giúp giải trí và phát triển kỹ năng."
   "true-label": "REFUTES"
   "predicted-label": "SUPPORTS"
   }

   The model tends to assign the label SUPPORTS because the *evidence* mentions negative consequences of excessive gaming. However, logically, the *evidence* actually refutes

the absolute nature of the *claim* by introducing a condition (moderate gaming) that leads to positive outcomes. This indicates that the model lacks sufficient capability to handle conditional expressions or logic involving degrees (quantification) in Vietnamese.

The two examples above highlight that most classification errors stem from subtle linguistic phenomena such as negation structures, implication, and the need for background knowledge. In future development phases, we plan to augment the training data with more examples of such challenging cases, while also incorporating logical reasoning steps or external knowledge sources to improve accuracy.

6.2 Visualized Result

After training and evaluating multiple models for the fact-checking task, MDeBERTa was selected as the best-performing model based on evaluation metrics such as F1-score, precision, and recall on the test set. To gain deeper insights into classification performance and the distribution of predicted labels, we visualized the model outputs using summary tables and various charts, as presented below:

claim	evidence	true_label	predicted_lab
Dốc Lết ở Ninh Hòa mình dạo này rác kinh khủng, chả ai dọn dẹp gì cả.	Một số du khách phản ánh tình trạng xả rác tại bãi biển Dốc Lết, gây mất mỹ quan và ô nhiễm.	SUPPORTS	SUPPORTS
học giỏi hay không thì quan trọng là môi trường học tập và sự nỗ lực của mỗi cá nhân	Môi trường học tập tích cực và sự nỗ lực cá nhân là yếu tố quan trọng quyết định thành công học tập	SUPPORTS	SUPPORTS
Trục 'kéo co' là một trò chơi dân gian chỉ có ở miền Bắc Việt Nam.	Kéo co là một trò chơi dân gian phổ biến và là nét văn hóa truyền thống ở nhiều vùng miền của Việt Nam, từ Bắc vào Nam, thường xuất hiện trong các lễ hội làng xã. (Nguồn: Văn hóa dân gian Việt Nam)	REFUTES	REFUTES
Chơi game cùng bạn bè vui lắm	Chơi game nhóm tạo sự gắn kết và mang lại niềm vui xã hội.	SUPPORTS	SUPPORTS
Ăn uống khoa học để khỏe đẹp	Chế độ ăn khoa học, cân bằng dinh dưỡng giúp duy trì sức khỏe và vóc dáng đẹp	SUPPORTS	SUPPORTS
Trường nào ở Hà Nội mà sinh viên nữ mặc áo dài đi học mỗi thứ 2 đầu tuần à mọi người? Hình như Đại học Sư Phạm Hà Nội phải không?	Trường Đại học Sư Phạm Hà Nội yêu cầu sinh viên nữ mặc áo dài vào mỗi sáng thứ Hai đầu tuần, theo truyền thống của trường.	SUPPORTS	SUPPORTS
tôi hay mua đồ online lắm, tiện lợi	Mua hàng online giúp tiết kiệm thời gian, đa dạng lựa chọn và thuận tiện cho người dùng	SUPPORTS	SUPPORTS
mỗi người có cảm nhận vị khác nhau nữa	Mỗi người có sự cảm nhận và ưu tiên vị giác khác nhau do yếu tố sinh học và thói quen.	SUPPORTS	SUPPORTS
Nghe nói Netflix sắp tăng giá gói Premium ở Việt Nam lên 300k/tháng.	Hiện tại Netflix chưa công bố kế hoạch tăng giá gói Premium tại Việt Nam, giá vẫn giữ mức dưới 260.000 đồng/tháng.	REFUTES	REFUTES
Uống một ly sữa ấm trước khi đi ngủ giúp dễ ngủ hơn	Sữa ấm giúp cơ thể thư giãn, dễ chịu hơn, nhờ đó hỗ trợ giấc ngủ đến nhanh hơn và sâu hơn.	SUPPORTS	SUPPORTS

Figure 6: Data visualization table

Figure 6 presents several data samples from the test set after being predicted by the **MDeBERTa** model. Direct observation of correct and incorrect cases helps qualitatively evaluate the model's effectiveness and identify typical classification errors.

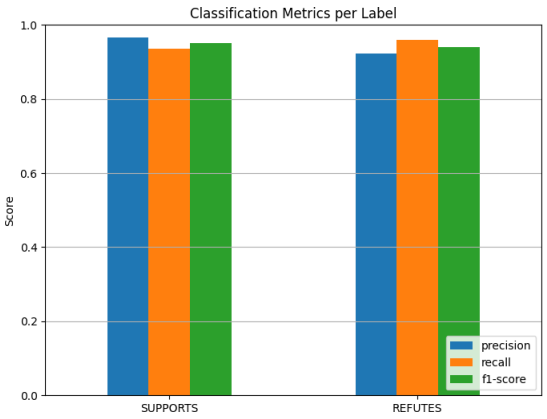

Figure 7: Classification metrics by label

Figure 7 illustrates the classification evaluation metrics for the two labels "SUPPORTS" and "REFUTES," including Precision, Recall, and F1-score. It can be observed that the model achieves relatively balanced performance between the two labels, with the "SUPPORTS" label having a slightly higher F1-score. This indicates that the model tends to classify truthful claims better while still maintaining reasonable performance on false claims.

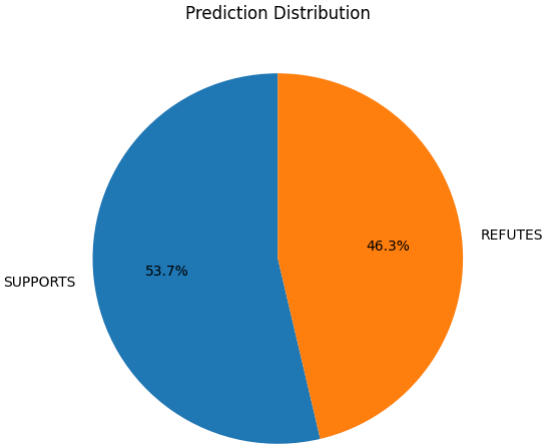

Figure 8: Pie chart of labels

The distribution of predicted labels by the model over the entire test set is shown in Figure 8. The distribution between the two labels, "SUPPORTS" and "REFUTES," is fairly balanced, indicating that the model does not exhibit label bias, and ensuring effective and fair recognition of both types of information.

7 Conclusion and Future Work

7.1 Conclusion

This study aims to improve the accuracy and efficiency of Vietnamese fact-checking systems in the

context of increasingly widespread misinformation on social media. We applied Machine Learning and Deep Learning techniques to develop and compare the performance of several models, including MDeBERTa, PhoBERT, viBERT, viT5, and DistilUSE + LR.

Evaluation was conducted on a dataset developed by our team, named ViFAiC. Experimental results show that the MDeBERTa model achieves the highest performance, with an F1-score of 94.6% and an accuracy of 94.6% on the test set.

This research has the potential to support the development of automated Vietnamese fact-checking systems, contributing to the mitigation of fake news and promoting a transparent and trustworthy social media environment.

7.2 Limitations

Although ViFAiC was developed with the goal of fact-checking within the context of real Vietnamese language usage, the majority of the current dataset (10,000 samples) is generated by large language models (LLMs), while only approximately 2,000 claims were directly collected from social media platforms such as Facebook.

The low proportion of social media data imposes certain limitations on the representativeness of everyday language, especially in cases involving ambiguous, unstructured, or colloquial expressions — features commonly found on digital platforms. Moreover, the automated content moderation processes employed by social media platforms complicate the collection of statements related to sensitive topics, thereby reducing the diversity and topical coverage of the dataset. This represents one of the key areas the research team intends to address in future development stages.

7.3 Future work

In the future, the research team plans to expand the collection of diverse claims from popular social media platforms such as Facebook, TikTok, and Threads, aiming to more accurately reflect the real information flows within the community.

Simultaneously, we will enhance the construction and verification of evidence from official and reliable sources, as well as leverage the data generation capabilities of large language models to produce supporting evidence more precisely and effectively.

Moreover, the team intends to develop more suitable preprocessing and pretraining techniques tailored for social media data — which often contains non-standard language, fragmented sentences, and diverse linguistic elements such as slang, emojis, hashtags, and accompanying multimedia formats. This will help improve the quality of inputs for fact-checking models.

Additionally, we plan to integrate multi-task learning and multimodal learning methods, enabling the models not only to rely on textual data but also to utilize images, videos, or audio related to claims, thereby enhancing the accuracy and generalizability of the models.

Finally, the research will extend practical applications by deploying automatic fact-checking models on social media platforms to evaluate their effectiveness in real-world environments. Concurrently, collaboration with experts in sociology and psychology will be pursued to develop community awareness programs about fake news and misinformation, contributing to building a more transparent and trustworthy online information environment for Vietnamese users.

References

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. ViFactCheck: A New Benchmark Dataset and Methods for Multi-domain News Fact-Checking in Vietnamese. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pages 308–316, Philadelphia, Pennsylvania, USA. Association for the Advancement of Artificial Intelligence.

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.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525. Association for Computational Linguistics.