Automatic Fake News Detection on Online Social Networks through Natural Language Inference and Logical Reasoning
(Phát hiện Tin giả trên Mạng xã hội Trực tuyến thông qua Suy luận Ngôn ngữ Tự nhiên và Suy luận Logic)

Le Quoc Bao[1]_2252065_Computer Science, Quan Thanh Tho[1]
[1] Office for International Study Programs, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam.
Corresponding Author: qttho@hcmut.edu.vn

Abstract

With the development of science and technology, social media has become extremely popular in daily life. Along with the strong increase in the number of users on social media, the number of posts has also increased extremely rapidly. Social media is now not only a place for socializing, exchanging, and entertainment, but also a place for people to easily, quickly, and conveniently update news. However, many individuals and organizations exploit social media to spread fake news for malicious purposes, especially regarding medical news since the outbreak of COVID-19, which has attracted a large number of social media users' attention. Receiving this fake news not only causes public confusion and social disorder but also damages the reputation of individuals, organizations, and businesses. Furthermore, fake medical news can have many health-related consequences, even endangering the lives of those who receive it. Therefore, the automatic detection of fake news on social media has become a topic of great interest. There have been many studies on this issue, mainly in the English language. In our work, we propose a model, **ExFAN**, to automatically detect fake news with explanations. The model is built based on *Transformer* models and formulates the problem as an *NLI* task (Natural Language Inference). Additionally, we also construct a fake news dataset, **ViFactOSNs**, based on posts on the OSNs (Online Social Networks) with context and evidence. Evaluation results on various datasets and different setups show that the model has achieved better results compared to baseline models and competes with other state-of-the-art models in the same research direction.

Keyword: Fake-News, COVID-19, Transformers, ExFAN, NLI, Semantic Similarity.

1. Introduction

With the development of the Internet and social media, fake news has become a widespread issue encompassing various types of deception, such as misleading reviews, fake online accounts, and harmful websites. According to VnExpress[i], automatic language generation models like ChatGPT have taken the creation of a slew of fake content to new heights, making it increasingly difficult to distinguish between true and false information.

Detecting fake news on social media poses new and challenging research issues. As AI-Generated fake news becomes increasingly sophisticated, many automatic fraud detection models have been proposed.

There are various approaches to the problem of fake news detection, but in general, it can be divided into 2 approaches [1]:

1. **Non-Interpretable Fake News Detection Method:** This approach relies on content, based on the writing style such as syntax, textual meaning; using simple strategies such as capturing punctuation, vocabulary, and the text's emotional tone, which serves as the foundation for this approach. Additionally, the contextual approach focuses on the social media aspects related to users (profiles, posts, comments,...).
2. **Interpretable Fake News Detection Method:** In this approach, the problem is modeled as a logical process, where external evidence is provided to verify the accuracy of the information. This model needs to help discover and integrate useful information from evidence for confirming news.

Automatic fake news detection is a practically significant problem and brings benefits to social media users. Therefore, within the scope of our research, we have tackled this topic to address the issue in Vietnamese news. In this work, we present our research efforts, the methods used in the fake news detection problem, and some experiments on explanatory fake news datasets.

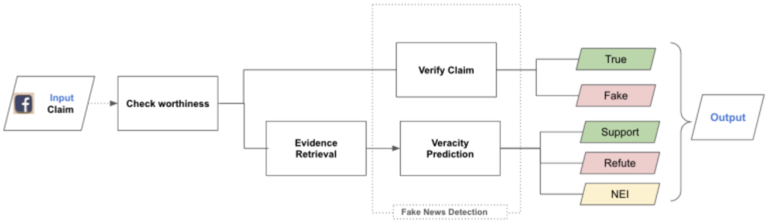
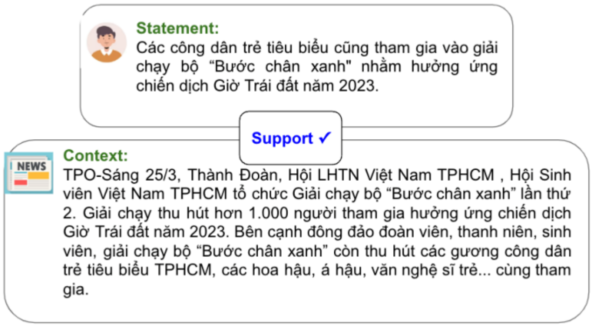


Figure 1: *Illustration of the fake news detection problem.*

Figure 1 describes the main tasks in fake news detection using text, including the following problems:

- **Check worthiness:** Identifying claims that need verification
- **Verify Claim:** Assessing the truthfulness of a claim based on its own features.
- **Evidence Retrieval:** Searching for supporting or refuting evidence for a claim.
- **Veracity Prediction:** Verifying the truthfulness of a claim based on the evidence collected in the **Evidence Retrieval** step.

However, in the scope of research and the implementation of this research, we will focus on **a problem of checking fake news using evidence**, specifically claims supported by news, articles collected from different sources, and categorizing the claim-evidence pairs as *SUPPORTS, REFUTES, or NO-INFO* (as in **Figure 2**). This limitation simplifies the training data setup but is still effectively applicable in practice (as it provides specific evidence for fact-checking).



Statement:
Các công dân trẻ tiêu biểu cũng tham gia vào giải chạy bộ "Bước chân xanh" nhằm hưởng ứng chiến dịch Giờ Trái đất năm 2023.

Support ✓

Context:
TPO-Sáng 25/3, Thành Đoàn, Hội LHTN Việt Nam TPHCM , Hội Sinh viên Việt Nam TPHCM tổ chức Giải chạy bộ "Bước chân xanh" lần thứ 2. Giải chạy thu hút hơn 1.000 người tham gia hưởng ứng chiến dịch Giờ Trái đất năm 2023. Bên cạnh đông đảo đoàn viên, thanh niên, sinh viên, giải chạy bộ "Bước chân xanh" còn thu hút các gương công dân trẻ tiêu biểu TPHCM, các hoa hậu, á hậu, văn nghệ sĩ trẻ... cùng tham gia.

Figure 2: *Illustrative example of a claim-evidence pair labeled as **SUPPORTS**.*

- **Input:** The inputs for the fake news detection problem include a claim *c*.
- **Output:** A list of evidence $\hat{\varepsilon}(e)$ and a label $y(c, e) \epsilon$ {SUPPORTS, REFUTES, NO-INFO}.

2. Methods and Implementation

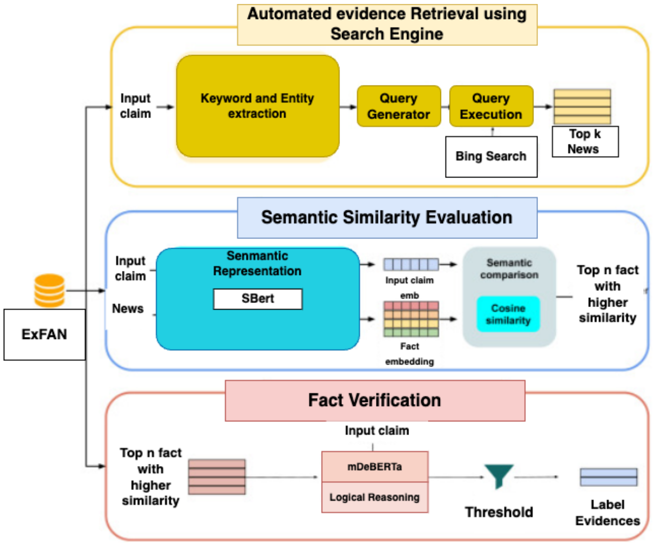


Figure 3: *Architecture of the proposed ExFAN model.*

Figure 3 illustrates the architecture of the proposed **ExFAN** model based on the **FacTer-Check** model [2]. The model consists of three main tasks: Automated evidence retrieval using search engine, Semantic Similarity Evaluation, and the final task of news classification. Along with this are the proposed changes compared to the original model, which include:

- In the Retrieval evidence engine task, changing the information source from fact-checking websites to a **Search Engine**.
- Changing the **semantic similarity** models from *Multilingual SBert* to *Vietnamese SBert*.
- Changing the **NLI** architecture from *XML-Roberta* to *mDebertaV3*.
- Adding **Logical Reasoning** into *Fact Verification* task.

2.1. Automated evidence retrieval using search engine

Figure 4 illustrates the architecture of the task for automated evidence retrieval from the Internet. An article posted by a user in a social media group on Facebook essentially contains information about the content of the article, represented by the Post content. After preprocessing and tokenization, the Post Content *P* yields a corresponding query.

Due to the specific nature of the Vietnamese language compared to English, for normalizing Unicode and accent marks in Vietnamese, we uses open source code from GitHub, with additional modifications for suitability.

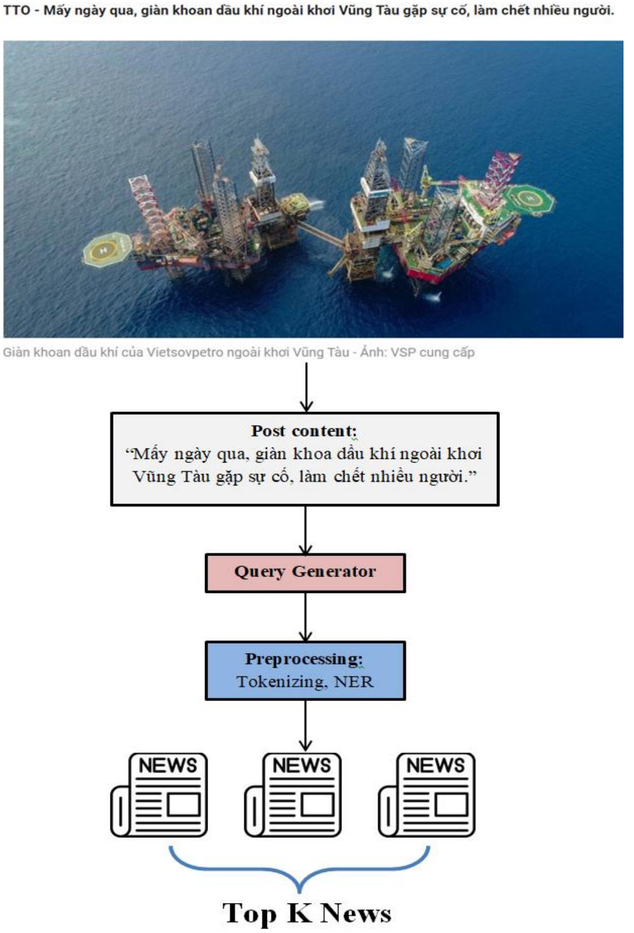

TTO - Mấy ngày qua, giàn khoan dầu khí ngoài khơi Vũng Tàu gặp sự cố, làm chết nhiều người.
Giàn khoan dầu khí của Vietsovpetro ngoài khơi Vũng Tàu - Ảnh: VSP cung cấp
Post content:
"Mấy ngày qua, giàn khoa dầu khí ngoài khơi Vũng Tàu gặp sự cố, làm chết nhiều người."
Query Generator
Preprocessing:
Tokenizing, NER
NEWS
NEWS
NEWS
Top K News

Figure 4: *Architecture of the* ***Automated evidence retrieval using search engine*** *task.*

For applying Sentence segmentation and Tokenization, there are several libraries supporting Vietnamese language preprocessing, such as pyvi, underthesea[ii], and more. We opt for the underthesea library due to its comprehensive development, fast processing speed, and regular updates. Underthesea is a toolkit for research and development in natural language processing for Vietnamese, launched in March 2017. Despite other good existing toolkits for Vietnamese such as vn.vitk, pyvi, there was still a lack of a comprehensive, opensource, easily installable and usable toolkit like the equivalent products for English, such as nltk, polyglot, spacy.

Additionally, for Vietnamese stop words; we will use an open-source repository on GitHub. This stop words repository contains 1942 common stop words in both spoken and written Vietnamese language.

For Vietnamese text, some preprocessing techniques differ from English, including:

- Normalizing VNM Unicode: Encoding all characters to standard Unicode.
- Normalizing VNM accent marks placing: Converting accent marks of words to standard pre-tone marks.
- Tokenization: Applying the word_tokenize function of underthesea to split strings into meaningful words, while also recognizing Named Entities. After this step, the processed string from the previous two steps is split into a list of individual words, compound words, or punctuation.
- Removing stopwords and punctuation: Examining each element in the list of tokens and only keeping it if it is not a stop word or punctuation. Applying a separate stop word dictionary for Vietnamese.

In the context of search control, several search engines, such as Google, Bing, ChatNoir, etc, are available. The search results are obtained from the returned HTML files, which contain the URLs of the search results. Among these, Google provides the fastest search results for a 20-word query, demonstrating highest efficiency. However, Google's anti-bot system easily detects and blocks multiple consecutive searches. In contrast, other search engines, allow multiple searches without triggering anti-bot measures. Bing, in particular, yields the best results for both Vietnamese and English queries and provides the fastest response times among the alternatives. Therefore, the decision was made to implement URL-based searches using the GET method, specifically through Bing.

2.2. Semantic similarity evaluation task

In this task, with the Vietnamese language dataset, we propose replaces the architecture of the *Multilingual SBert* model with the *Vietnamese Sentence* model for Vietnamese by Long and colleagues[iii]. This shift is due to the Vietnamese SBert model achieving the highest accuracy **(95.33%)** and F1-score **(95.42%)** in the Vietnamese Paraphrasing task. These results outperform many recent models for the task of paraphrasing on the VnPara Corpus. This demonstrates the effectiveness of the Vietnamese SBert model with strong support from PhoBERT, a pre-trained language model for Vietnamese.

Based on the original SBERT model, which includes two layers: transformer and pooling, the architecture of the Vietnamese SBert model is depicted in **Figure 5.**

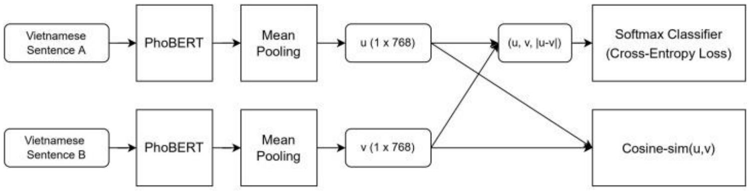


Figure 5: *Architecture of the Vietnamese SBERT model*

To train the model, the authors use pairs of labeled Vietnamese sentences corresponding to each type of relevant linguistic data (including NLI and STS Datasets).

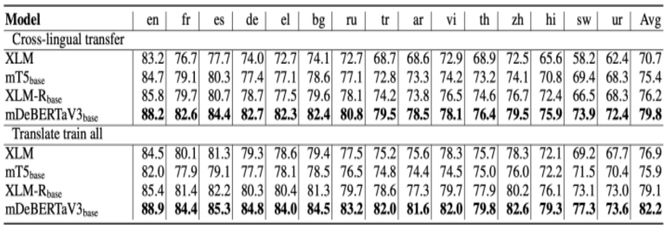
Subsequently, we configure the transformer layer using the PhoBert-base [3] model to extract Word Embeddings for individual words or compound words in Vietnamese. Since the release of PhoBERT, it has been the most modern and best performing transformer model for Word Embedding tasks in Vietnamese. Therefore, the authors use PhoBERT as a component in the SBERT model [4].

The Embedding Vector representing each sentence is then passed to the pooling layer, using the *mean method* due to its proven superior performance, as demonstrated in the paper by *Reimers and colleagues* in 2019 [4]. This reduces the number of features in the representation vector of each sentence down to a fixed length embedding vector (*768*).

Finally, the sentence transformer model receives the output from the two layers above and adds a Siamese Neural Network [5] and Triple Network to update and extract semantic attributes to represent a sentence. To fine-tune the model with the NLI dataset, the authors use a *3-way Softmax Function* as the *Classification Objective Function*, corresponding to the 3 classification labels mentioned earlier. Then, to fine-tune the model with the STS dataset, the authors use a *Regression Objective Function* employing the *cosine similarity* distance function for a sentence pair and use *mean-squared-error* (MSE) as the objective function.

2.3. Fact Verification Task

We propose to change the NLI model architecture from **XML-Roberta** to **mDebertaV3** compared to the reference model. The DeBERTa-v3 model [6] is an improvement of BERT with an enhanced training process based on the Attention mechanism.



Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Cross-lingual transfer																
XLM	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
$\mathrm{mT5}_{\mathrm{base}}$	84.7	79.1	80.3	77.4	77.1	78.6	77.1	72.8	73.3	74.2	73.2	74.1	70.8	69.4	68.3	75.4
$\mathrm{XLM\text{-}R}_{\mathrm{base}}$	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
$\mathrm{mDeBERTaV3}_{\mathrm{base}}$	**88.2**	**82.6**	**84.4**	**82.7**	**82.3**	**82.4**	**80.8**	**79.5**	**78.5**	**78.1**	**76.4**	**79.5**	**75.9**	**73.9**	**72.4**	**79.8**
Translate train all																
XLM	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
$\mathrm{mT5}_{\mathrm{base}}$	82.0	77.9	79.1	77.7	78.1	78.5	76.5	74.8	74.4	74.5	75.0	76.0	72.2	71.5	70.4	75.9
$\mathrm{XLM\text{-}R}_{\mathrm{base}}$	85.4	81.4	82.2	80.3	80.4	81.3	79.7	78.6	77.3	79.7	77.9	80.2	76.1	73.1	73.0	79.1
$\mathrm{mDeBERTaV3}_{\mathrm{base}}$	**88.9**	**84.4**	**85.3**	**84.8**	**84.0**	**84.5**	**83.2**	**82.0**	**81.6**	**82.0**	**79.8**	**82.6**	**79.3**	**77.3**	**73.6**	**82.2**

Figure 6: *Results of mDebertaV3 compare to other multilinguals models on XNLI test set under the cross-lingual transfer and the translate-train-all settings.*

Specifically, our goal is to enhance the fact-checking and misinformation detection process by using the DebertaV3 model for Natural Language Inference (NLI). The DebertaV3 model, with its deep understanding and high accuracy, is chosen to perform the inference task between pairs of statements. The transition from the Roberta model to DebertaV3 is driven by its superior performance and applicability in multilingual language contexts. The DebertaV3 model has achieved significantly better results than other models such as BERT and BioBERT on multilingual datasets for fake news and NLI.

The DebertaV3 model provides profound semantic analysis and understanding of sentence grammar structure. Combined with the use of multilingual NLI data, we expect the model to yield superior results and improve its applicability across various contexts.

Proposal: Addition of Logical Reasoning

In the field of natural language processing, integrating Logical Reasoning (LR) into Natural Language Inference (NLI) is considered an important development direction to enhance text understanding and inference. However, the current major challenge is how to create a stronger NLI system, especially through integrating components such as Named Entity Recognition (NER), parser, and semantic awareness to support the Logical Reasoning process.

Key Components and Their Functions:

- **Named Entity Recognition (NER):** Identifies and classifies significant entities in text.
- **Parser:** Analyzes grammatical structure of sentences.
- **Semantic Awareness:** Enhances understanding of word meanings and context.

Type	Example
Number	**Claim:** Năm nay, khoảng **200.000** học sinh ở TP.HCM tốt nghiệp THCS. **Evidence:** Năm nay,khoảng **100.000** học sinh ở TP HCM tốt nghiệp THCS.
Proper Noun	**Claim:** Một số người khuyên Nhi Pham theo Công nghệ thực phẩm, thay vì Công nghệ dệt may, nhưng em ghét **Toán**. **Evidence:** Một số người khuyên em theo Công nghệ thực phẩm, thay vì Công nghệ dệt may, nhưng em ghét **Hóa**.
Unit	**Claim:** Mẫu vật lớn nhất dài tầm 1 **m**. **Evidence:** Mẫu vật lớn nhất dài tầm 1 **cm**.
Acronym	**Claim:** Kế hoạch thi, xét tuyển cụ thể **đã** được thông báo **trước**. **Evidence:** Kế hoạch thi, xét tuyển cụ thể **sẽ** được thông báo **sau**.
Negation	**Claim:** Ở giai đoạn đầu tiên, một đội nhân công địa phương **không được** tập huấn về quá trình in. **Evidence:** Ở giai đoạn đầu tiên, một đội nhân công địa phương **được** tập huấn về quá trình in.

Table 1: *Examples of apply logical reasoning for detect fake news with Claim (C) – Evidence (E) pairs.*

The goal of using NER and parser is to identify and analyze syntactic entities, providing a foundation for logical relationships within the text. Additionally, the use of semantic awareness improves logical inference by better understanding context and meaning. Combining NER, parsers, and semantic awareness promises significant progress in NLI capabilities. This integration improves text comprehension and inference, offering practical benefits across various real-world applications. An example of using logical reasoning is illustrated in **Table 1**.

3. Results and Discussion

3.1. Overview of the Datasets

To consider a dataset as a truth verification dataset, it needs to provide claims, evidence (text or sentences), and final verification labels. Such a dataset allows for both the tasks of evidence retrieval and claim verification. This is important because the ultimate goal of many automatic truth verification systems is to mimic the work of experts, where both evidence search and conclusion drawing constitute the process. The dataset **DS01-ISE-ICHEVE (Information Checking and Verification)** is a dataset in the UIT Data Science Challenge in 2023, and finally, the misinformation dataset **ViFactOSNs** has evidence that the learners built from the original **ReINTEL**[iv] dataset. Details are as follows:

Language	Dataset	Claim	Claim Origin	Evidence Source	Domain
Vietnamese	**DS01-ISE-ICHEVE**	26,576	News Articles	News Articles	Multi-Domain
Vietnamese	**ViFactOSNs (ours)**	5,388	Social Media	Search Engine	Health

Table 2: *Relevant information about the fact-checking datasets*

3.1. Evaluation Method

For the specific problem of detecting fake news based on evidence, several metrics are used to evaluate the model's performance. In our work, we use 5 metrics to evaluate the experimental results on different datasets: *Precision, Recall, F1 Score, Accuracy,* and *Strict Accuracy.*

Condition	Interpretation
True Positive (TP)	Abnormal points correctly detected as actual abnormal points
True Negative (TN)	Non-abnormal points correctly detected.
False Positive (FP)	Non-abnormal points incorrectly predicted as abnormal
False Negative (FN)	Abnormal points not detected.

Table 3: *Conditions and explanations of values in the confusion matrix*

The confusion matrix (or error matrix) is a table designed to visualize the results of an algorithm. It is particularly useful when the dataset is imbalanced, as accuracy alone does not adequately reflect the algorithm's true performance. To evaluate the detection function, the conditions of the confusion matrix are detailed in **Table 3.**

- **Accuracy:** The most intuitive measure, it is simply the ratio of the correctly predicted observations to the total observations. Although widely used, accuracy is not the most appropriate measure in some cases, especially when classes in the dataset are imbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Strict Accuracy**

  Let v and v' be the sample verdict and predicted verdict ($v, v' \in$ {SUPPORTED, REFUTED, NEI}).

  Let e and e' be the sample and predicted evidence.

$$StrAcc = \delta(v, v') \times \delta(e, e')$$

  Where δ is the Kronecker symbol with $\delta(x, y) = 1 \Leftrightarrow x = y$, and $\delta(x, y) = 0 \Leftrightarrow x \neq y$.

- **F1 Macro score**

  The F1 score considers both precision and recall to compute the model's performance. Mathematically, it is the harmonic mean of the model's precision and recall, computed as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

  This F1 score is computed across different classes, taking the average of the F1 scores of all classes to obtain the average.

3.2. Baseline Models

We compare **ExFAN** model with several **SOTA** models in the scientific fact verification task related to specific scientific and medical datasets for Vietnamese Datasets:

- **XLM-RoBERTa** [7] is the latest multilingual transformer-based language representation model. This model is an extended version of the BERT model trained on a large-scale text corpus with a hundred languages. Largescale empirical experiments conducted by the authors have shown that the *XLM-RoBERTa* model outperforms traditional *mBERT* in a range of multilingual text classification tasks.
- The pre-trained **PhoBERT** model published by VinAI [3] for feature extraction specifically for Vietnamese.

3.3. Experimental Results and Discussion

Rank	Model	ViFactCheckOSNs		
		Precision	Recall	F1 Score
1	XLM-RoBERTa	66	60	62.8
2	PhoBERT	**75**	68	71.3
3	**ExFAN**	73	**76**	**74.5**

Table 4: *Results on the **ViFactCheckOSNs** Dataset*

Table 4 presents the experimental evaluation results of the **ExFAN** model and the baseline models on the **ViFactCheckOSNs** dataset. Overall, the results of the **ExFAN** model show an improvement across all 3 metrics compared to modern text classification models today, indicating that the approach to natural language inference and the use of the *DebertaV3* model have improved the performance of the fake news detection task with evidence. Specifically, **ExFAN** model also outperforms the feature extraction model specifically for Vietnamese, **PhoBERT**, by 3%, and outperforms the multilingual model **XLM-RoBERTa** by 10%.

Rank	Model	UIT Data Science Challenge 2023		
		Strict Acc	Acc	Acc@1
1	UIT9	79.1197	84.4389	79.5578
2	**ExFAN**	77.8725	83.7102	79.9666
3	ViNSV	76.3343	81.6716	78.1134

Table 5: *UIT Data Science Challenge 2023 Dataset results*

Table 5 presents the experimental evaluation results of the **ExFAN** model compared to other teams in the competition **(Top 3)**. The results of the **ExFAN** model are quite competitive with the models that ranked first. Additionally, the model has high generalization ability and can make predictions without requiring a training stage, whereas using a solution like basic machine learning models such as **SVM** may be difficult to implement. This is in line with the goal of our work, which is to detect fake news on social media.

4. Conclusions

In this paper, we have introduced and investigated the creation of the **ViFactOSNs** dataset, a significant contribution to the field of fake news detection on Vietnamese social media. Through exploration of the Fake News Detection problem, our development of **ExFAN** fake news detection model represents a substantial step forward, aiding in evidence retrieval, verifying information, and presenting conclusions to help end-users check fake news simply and convincingly.

However, our efforts encountered several limitations, including challenges related to searching for relevant information within excessively long social media posts, resource constraints during the experimentation process, and time limitations preventing real-world implementation and validation of the model's effectiveness.

Moving forward, we propose several avenues for future research and development, including further exploration of recent models and large-scale data models, real-world implementation of the proposed model to monitor its stability and effectiveness, and continued data collection to build a robust fake news detection system in Vietnamese. Our work sets the stage for continued progress in the automatic detection of fake news on online social networks, providing a framework for future research and development in this critical area of study.

5. References

[1] Qiang Sheng, Xueyao Zhang, Juan Cao and Lei Zhong, "Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning," *CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management,* p. 1640–1650, October 2021.

[2] A. Martín, "FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference," *Knowledge-Based Systems,* vol. 251, p. 109265, 2022.

[3] Dat Quoc Nguyen and Anh Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese.," *Findings of the Association for Computational Linguistics: EMNLP 2020,* p. 1037–1042, 2020.

[4] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Empirical Methods in Natural Language Processing (EMNLP),* 2019.

[5] Sumit Chopra, Raia Hadsell and Yann LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face," *Conference on Computer Vision and Pattern Recognition (CVPR),* 2005.

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao and Weizhu Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *ICLR 2021 - The Ninth International Conference on Learning Representations,* 2020.

[7] Guillaume Lample, Alexis Conneau, "Cross-lingual Language Model Pretraining," *Advances in Neural Information Processing Systems 32,* p. 7059—7069, 2019.

[i] Mối nguy AI "siêu lan truyền" thông tin sai lệch

[ii] Under The Sea Toolkit: *https://github.com/undertheseanlp*

[iii] Vietnamese SBERT model: *https://huggingface.co/keepitreal/vietnamese-sbert*

[iv] VSPL_2020_ReINTEL Dataset: *https://huggingface.co/ datasets/truongpdd/vslp_2020_ReIntelt*