Utilizing Transformer Models To Detect Vietnamese Fake News on Social Media Platforms

Anh Tuan Huynh[1], and Phuoc Tran[2]
[1] Student from the Faculty of Information Technology, Ton Duc Thang University,
Ho Chi Minh City, Vietnam
[e-mail: huynhanhtuan02.tv@gmail.com]
[2] Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology,
Ton Duc Thang University,
Ho Chi Minh City, Vietnam
[e-mail: tranthanhphuoc@tdtu.edu.vn]
*Corresponding author: Gildong Hong

Abstract

The spread of fake news on social media has become a serious issue, leading to misinformation and causing harm to society. This project aims to develop a system for analyzing and classifying Vietnamese fake news using transformer models, with a particular focus on PhoBERT - a version of BERT optimized for Vietnamese. To address this issue, we collected a dataset consisting of Vietnamese posts on the Facebook social media platform and several articles from Vietnamese news sources, covering topics such as lifestyle news, current affairs, and politics. However, there are still challenges due to data imbalance between the number of true and false news. The posts were manually labeled as true or false, then underwent data preprocessing and were trained using transformer models and PhoBERT for Vietnamese. We also incorporated the TF-IDF data preprocessing technique to optimize the model's performance. To evaluate the performance of the models, we used various evaluation metrics such as Accuracy, Precision, Recall, F1 Score, and AUC. Our results indicate that PhoBERT outperforms other transformer models in detecting Vietnamese fake news, achieving high accuracy and reliability. This report outlines the background, objectives, methodology, and future research directions, providing a comprehensive overview of the project and its contributions to the field of fake news detection.

Keywords: Fake News Detection, PhoBERT, Social Media Analysis, Transformer Models, Vietnamese Fake News.

1. Introduction

In the context of global modernization, social media platforms are becoming increasingly popular, which comes with positive and negative impacts. In particular, the rapid spread of fake news on social networks has emerged as a serious social problem, when false information is spread causing many misunderstandings, even conflicts globally.

In the case of Vietnam, fake news has frequently caused public uproar, typically fake news related to epidemics, traffic accidents, incorrect knowledge in daily life, and even politically subversive content. These types of fake news often spread quickly in the community, causing confusion in public opinion and affecting people's lives. Therefore, researching and detecting fake news is a necessary task to support and maintain social stability. For that reason, we chose this topic as our research object.

In the past few years, deep learning has been recognized as a powerful tool in the field of artificial intelligence, especially in natural language processing (NLP). However, traditional deep learning models often rely on sequential data processing [1], which can be limiting when faced with complex language tasks. Then, the introduction of a new architecture, Transformers, revolutionized NLP by using attention mechanisms, allowing for more efficient processing of context and relationships in text [2]. These advantages help Transformers perform better in understanding the language and context of text, thereby solving text classification tasks such as detecting fake news.

Throughout this research, we focus on leveraging the Transformer models BERT and variants to detect fake news, especially using PhoBERT - a variant designed specifically for the Vietnamese language [3]. Our goal is to develop an effective system to identify fake news on social media platforms, especially Facebook - the most popular social media platform in Vietnam.

By harnessing the power of PhoBERT, we aim to improve accuracy and efficiency in detecting Vietnamese fake news. However, we are facing major challenges due to the lack of large-scale datasets containing both real and fake news in Vietnamese. We have collected data from official Facebook pages of press agencies in Vietnam and fake data sources from impostor pages, anti-establishment sources, and tabloid newspaper pages, including many fields from social life to politics. To do this work, we used several different tools, including Selenium to collect data, followed by data processing through cleaning and encoding.

The structure of the remainder of this paper is as follows: In Part 2, we review related works on transformer models and fake news detection, focusing on methods and models applied to research in Vietnam. Next, in Part 3, we present details of the proposed method, including the overall model and specific steps for system development. Part 4 focuses on the experimental setup, dataset description, results, and discussion. Finally, Part 5 summarizes our findings and suggests future research directions.

2. Background

2.1 Related work

Detecting fake news has become a significant area of research due to the increasing presence of misinformation globally. Several studies have explored various approaches to address this challenge.

The journey of detecting fake news has advanced significantly with the development of Transformer models. Vaswani et al. (2017) [4] introduced the Transformer architecture, which

uses a self-attention mechanism to efficiently process sequential data, thereby laying the foundation for modern NLP.

Since then, many Transformer models have been created to perform natural language processing tasks. In 2018, Devlin et al. [5] introduced BERT (Bidirectional Encoder Representations from Transformers), a model with bidirectional attention capabilities, helping the model better understand the context of words in sentences. Based on the foundation of BERT, Liu et al. (2019) [6] developed RoBERTa, improving training efficiency and performance on NLP tests. In addition, Sanh et al. (2019) [7] introduced DistilBERT, a more compact and faster version of BERT, suitable for applications that need fast response.

In the context of fake news detection, Agarwal et al. (2021) [8] used a Bi-LSTM layer with an attention function to classify English news based on context. Monti et al. (2019) [9] studied graph neural networks, using a four-layer Graph CNN to predict texts by combining information about user activities and articles. Meanwhile, Qi et al. (2019) [10] emphasized the importance of image content, presenting a multi-domain neural network using CNN and RNN models to analyze image features, helping to distinguish between fake news and real news.

In the Vietnamese context, Nguyen et al. (2020) [3] developed PhoBERT, a Transformer model pre-trained on a large Vietnamese text set. This has created a huge step forward for natural language processing (NLP) tasks in Vietnamese. Their results show that PhoBERT consistently outperforms the recent leading pre-trained multilingual model XLM-R. PhoBERT has helped improve performance on many Vietnamese-specific NLP tasks, including word classification, dependency analysis, name entity recognition, and semantic inference.

Recently, many studies have focused on using PhoBERT and other deep learning techniques to detect fake news in Vietnamese. One of the notable studies is that of Cao Nguyen Minh Hieu et al. [11] in the ReINTEL 2020 Competition. They developed a combinatorial model that combines PhoBERT embedding with time metrics and community interactions such as the number of shares, likes, and comments. Their StackNet model achieved an AUC score of 0.9521, topping ReINTEL's rankings.

Ngoc Dong Pham et al. (2021) [12] proposed a method that combines PhoBERT with TF-IDF to generate word embedding and uses CNN for feature extraction. This method achieved an AUC score of 0.9538. However, the reliance on the ReINTEL dataset may limit the diversity of the results. Cam Van Nguyen Thi et al. (2022) [13] introduced v3MFND, a deep multimedia multi-domain fake news detection model that integrates text, photos, and videos to improve accuracy, but the complexity of the model may affect its ability to real-time applicability. Khoa Dang Pham et al. (2023) [14] developed the vELECTRA model [15], using prefabricated features and achieving an AUC score of 0.9575 on the ReINTEL data set. However, reliance on these features can make it difficult to adapt to other situations. Meanwhile, Vo Trung Hung et al. (2022) [16] applied CNN and RNN models to classify news into four different groups, achieving an accuracy rate of 85%. Even so, the size of their dataset may reduce the generalizability of their results.

These studies show that Transformer models, especially PhoBERT, are very effective in detecting fake news in Vietnamese. They also highlight that combining text data with images, videos, and metadata can improve performance. However, there are still major challenges such as dataset size, diversity, and computational complexity that future research needs to address.

2.2 Theoretical basis

To effectively implement the project on detecting Vietnamese fake news on social media platforms using Transformer models, particularly PhoBERT and other BERT variants, it is essential to have a solid understanding of the following foundational knowledge:

2.2.1 Natural Language Processing (NLP)

Natural language processing (NLP) is a field in machine learning technology that allows computers to understand, interpret, and interact with human language [1]. NLP plays an important role in helping computers process and analyze text using machine learning and deep learning techniques. The main tasks in NLP include syntax analysis, semantic analysis, named entity recognition, and text classification.

In the text classification problem, natural language processing (NLP) will extract information from the text, process semantics, and represent the text as features that can be fed into machine learning models. or deep learning for classification. For example, language processing methods such as bag of words, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings help convert text into digital form. Then, machine learning models such as Naive Bayes, and SVM (Support Vector Machine),... can be trained to classify text into categories such as positive or negative, spam or not spam, and real or fake news.

2.2.2 Transformer Model

The Transformer model has truly created a breakthrough in the field of natural language processing, introduced by Vaswani et al. (2017) in the paper "Attention Is All You Need" [4]. Transformer stands out with its self-attention architecture, capable of understanding the relationship between words in a sentence without having to follow the sequential order like previous models, such as RNN or LSTM.

The Transformer model consists of two main components: the Encoder and the Decoder:

- Encoder: The Encoder takes in a sequence of words and converts them into semantic vectors. Each encoder is made up of several stacked layers, with two key components in each layer: the self-attention mechanism and the feedforward neural network. The self-attention mechanism helps the model focus on the important words in the sequence while filtering out the less relevant ones. After that, the feedforward neural network processes these attention-weighted vectors to generate deeper semantic representations.
- Decoder: The decoder works similarly to the encoder but has a few additional features. It uses self-attention to focus on the target input it's processing. Plus, it uses cross-attention to connect with the encoder's output. This setup allows the decoder to create meaningful representations based on both the original input sequence and the output sequence it has already generated.

The collaboration between the Encoder and Decoder allows the Transformer to process language tasks such as machine translation, text summarization, text generation, and text classification with flexibility and efficiency.

2.2.3 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained language model trained to understand the context of words in both directions (left to right and right to left) within a sentence [5]. It is trained through two main tasks:

- Masked Language Modeling (MLM): In this task, some words in the sentence are

replaced with the [MASK] tokens, and BERT has to guess the hidden words based on the surrounding words. This helps the model learn to understand the semantics of words in sentences without complete information, thereby improving its ability to grasp the meaning of words in many different situations.

- Next Sentence Prediction (NSP): This task requires BERT to predict whether a sentence is a continuation of a previous sentence, helping it improve the model's ability to understand relationships between sentences, this is very important in processing long and complex text.

BERT has demonstrated impressive performance in many language processing problems such as text classification, entity recognition, and question answering.

2.2.4 RoBERTa (A Robustly Optimized BERT Pretraining Approach)

RoBERTa is an improved version of BERT, designed to improve model performance by changing some methods such as skipping the Next Sentence Prediction (NSP) task. Instead, RoBERTa focuses on the task of Latent Language Modeling (MLM) and uses a much larger dataset than BERT for training [6]. As a result, RoBERTa has good performance in NLP tasks such as text classification and question answering, because of improvements in both data and training.

2.2.5 PhoBERT

PhoBERT is a special variant of BERT specifically trained on Vietnamese text data [3], helping the model better understand and capture the characteristics of Vietnamese. PhoBERT also builds on RoBERTa's improvements (removing the Next Sentence Prediction (NSP) task and focusing only on Masked Language Modeling (MLM)).

PhoBERT was trained on about 20GB of data, including about 1GB from Vietnamese Wikipedia and 19GB from Vietnamese news. In particular, before feeding data into the BPE encoder, PhoBERT used RDRSegmenter from VnCoreNLP to separate words, helping to improve the accuracy of language processing.

Because of training and fine-tuning on a large and diverse Vietnamese dataset, PhoBERT is better equipped to handle the complex structures of the Vietnamese language compared to BERT and RoBERTa.

2.2.6 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF (Term Inverse Frequency) is a popular technique in Natural Language Processing (NLP) and text mining [17], [18]. It helps evaluate the importance of a word in a document, based on both the frequency of that word in the document and the frequency of the word in the entire set of documents. In other words, TF-IDF allows us to determine which words are more prominent in a document compared to other documents in the same set.

In this study, we use TF-IDF as a preprocessing step to convert text into feature vectors. These vectors can then be combined with Transformer models to improve the ability to classify real and fake news. TF-IDF helps the model focus on important keywords and minimize the impact of common words that carry little information during model training.

3. Proposed methods

3.1 The designed system

Our system can be divided into four main stages , which are generally shown in **Fig. 1**: (1)

Data Collection, (2) Data Processing, (3) Model Training, and (4) Model Evaluation.

- Data Collection: In the first stage, our team collected data from Facebook posts, including both mainstream news sites and sites that frequently post false information on topics such as current affairs, lifestyle, and politics. We collect details like author, post content, post link, and also comments. This stage is very important because the collected data set will greatly affect the results of the research.
- Data Processing: Collected data will go through a series of pre-processing steps including cleaning, text normalization and the most important step is labeling articles with real or fake labels, which we do manually. After preprocessing, the data will be divided into training sets and test sets and ready for model training.
- Model Training: In this stage, we use the processed data to train the Transformer models: BERT, RoBERTa, and PhoBERT. We apply different training techniques to each model to optimize performance, including hyperparameter tuning, using techniques such as cross-validation. After training, we compare the results of the three models to evaluate their effectiveness in detecting fake news.
- Model Evaluation: The final stage involves evaluating the performance of the trained model. We use a separate test dataset to evaluate the model's accuracy, class precision, recall, and F1 score. Based on the evaluation results, we can further refine the model or adjust preprocessing techniques to enhance performance.

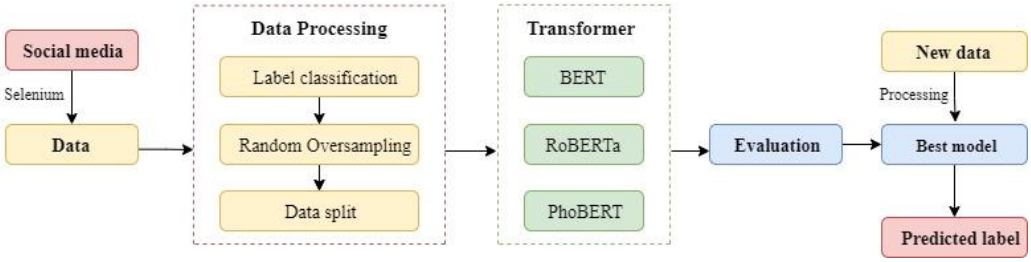

Fig. 1. General model of the system

3.2 Data collection

Because datasets on Vietnamese news and posts on social networks are limited and existing information may no longer be relevant to the current context. So we decided to collect our own data for research and hope to contribute to data resources to support future research.

We manually selected the posts. For real news, we identified official Vietnamese news pages on Facebook, including major media channels, government announcement pages, and other reputable sources. These are reliable sources for gathering authentic news. To collect fake news, we turned to tabloid pages and Facebook groups that regularly post sensationalist news and propagate false information about politics and society.

After selecting the necessary data sources, we used Selenium to automatically collect data, simulate user actions such as navigating the website, and extract data. Finally, we collected two datasets: one for authentic news and one for fake news, as shown in **Fig. 2**.

However, the data collection process faced several challenges, including limited collection time and difficulties in finding fake news sources due to some articles being removed after being reported. As a result, there is a discrepancy in the number of real and fake news items in our data.

date	author_id	content	label	link	comment_list
29/07/2024 13:45	https://www.facebook	Vụ xe bán tải cố vượt rào chắn, bị tàu hỏa tông ở Đ	0	https://www.fật" }, { "comment_id": "c36", "author": "Trần Phúc Hậu",	
30/07/2024 23:58	https://www.facebook	TPHCM: Hơn 4.600 ca mắc sốt xuất huyết, nhiều 'đ	0	https://www.fuyễn", "content": "Ánh Tây coi chừng bối nhen" }, { "comr	
30/07/2024 22:59	https://www.facebook	Nóng: Ngộ độc hàng loạt tại trụ sở công ty mẹ TikT	0	https://www.for": "Lê Nguyễn Bảo Thư", "content": "Nhii Mai Kim Ngân Hồng	
31/07/2024 14:50	https://www.facebook	Ngày mai: Giá xăng trong nước có thể giảm lần thứ	0	https://www.făng nào" }, { "comment_id": "c6", "author": "Lắm Chuyện	
31/07/2024 12:30	https://www.facebook	Pin dự phòng của hành khách bốc cháy tại nhà ga s	0	https://www.fỉm quá" }, { "comment_id": "c20", "author": "Thang Vo",	
31/07/2024 10:50	https://www.facebook	Thương tâm quá: Trong lúc chờ nhau trên xe máy v	0	https://www.fl": "c14", "author": "Vũ Hà", "content": "Nam mô a di đà phật	
30/07/2024 22:50	https://www.facebook	THƯƠNG TÂM HÀ GIANG: ĐẤT ĐÁ LĂN TỪ TALUY D	0	https://www.fuy", "content": "A di đà Phật" }, { "comment_id": "c9",	
30/07/2024 22:30	https://www.facebook	NỮ TÀI XẾ Ô TÔ ĐẠP NHẦM CHÂN GA GÂY TNGT LI	0	https://www.fạy mẹ trẻ ..may k chết ng ..." }, { "comment_id": "c12", "au	

Fig. 2. An example of the structure and content of some data

3.3 Data processing

We performed data cleaning through the following steps: removing empty, invalid, or duplicate entries, converting all text to lowercase, and eliminating special characters and URLs. Then we select the data fields that will be used and the remaining data will be shown as shown in **Fig. 3**.

	content	label
0	vụ xe bán tải cố vượt rào chắn bị tàu hỏa tông...	0
1	tphcm hơn 4600 ca mắc sốt xuất huyết nhiều điể...	0
2	nóng ngộ độc hàng loạt tại trụ sở công ty mẹ t...	0
3	ngày mai giá xăng trong nước có thể giảm lần t...	0
4	pin dự phòng của hành khách bốc cháy tại nhà g...	0

Fig. 3. Data after being cleaned and using information fields selected

At this stage, we are focusing solely on the content of the posts and their classification labels, but we plan to extend our research to include analysis of comments in the future.

It is clear that the number of fake news samples is significantly lower than that of real news (shown in the chart in **Fig. 4**), which could lead to bias in model training and inaccurate results. To address this issue, we have implemented two solutions:

- Incorporating Additional Data from VFND: We have added articles from the VFND dataset, described in the thesis of Ho Quang Thanh, "VNFD - Vietnamese Fake News Datasets" [19]. However, because this data set was collected from 2019-2020, we only selected news that has not changed over time, such as scientific knowledge that has been proven wrong, and news that has not changed over time. news about superstition, or deviant lifestyle. This additional data represents no more than 20% of the total fake news we currently collect.
- Using Random Oversampling Technique: We used the Random Oversampling technique from the "imbalanced-learn" library to rebalance the data. This is quite an effective way, by increasing the number of samples of labels with less quantity. This technique randomly copies existing samples of labels less than numbers until the number of samples of labels is balanced.

Fig. 4 and **Fig. 5** respectively show the ratio of two data labels before and after implementing the above two solutions for balance (label 0 represents real news and label 1 represents fake news). Balancing the labels helps prevent the model from being biased toward the majority class, improving accuracy for both labels, and ensuring that evaluation metrics such as precision, recall, and F1-score accurately reflect the model's true performance. This also enhances the model's ability to detect significant patterns, leading to improved AUC values.

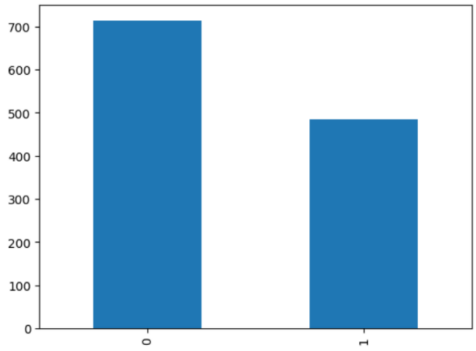

Fig. 4. Sample count for the two labels after collection

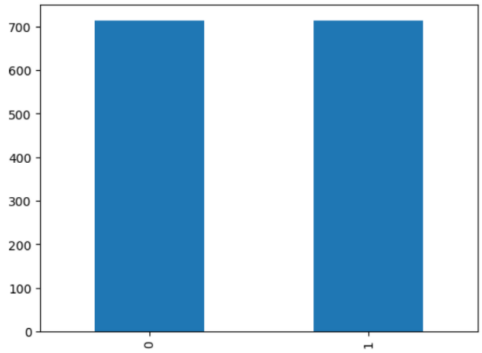

Fig. 5. Sample count for the two labels after processing

4. Experimental results

4.1 Corpus

After completing the data processing steps, including cleaning and balancing the data as discussed in sections 3.2 and 3.3, we obtained a dataset comprising social media posts and news from Vietnamese sources. The dataset contains over 1,400 samples, including both real and fake news across various domains. We then randomly split the dataset into a training set and a test set with an 80/20 ratio, resulting in 1,124 samples for training and 282 samples for testing. This approach allows the model to learn patterns effectively, increasing the likelihood of better performance on new, unseen data.

4.2 Corpus

The classification results will be evaluated using the following metrics:

- Accuracy: The ratio of correctly predicted samples to the total number of samples in the test dataset. It reflects the overall performance of the model, although it may not reflect accurately when the data is imbalanced. For instance, if a dataset has a significantly higher number of positive samples than negatives, then a model that predicts all outcomes as positive could still achieve high accuracy.
- Precision: The ratio of true positive samples to the total number of samples classified as positive by the model. It indicates the accuracy of the positive predictions.
- Recall: The ratio of correctly predicted positive samples to the total number of actual positive samples. A high recall means that the model misses few actual positive cases, demonstrating the model's ability to identify all positive cases.
- F1 Score: The harmonic mean of Precision and Recall, represents a balance of these two metrics. It is particularly useful in cases of imbalanced data. A higher F1 score shows that both precision and recall are high, and the model is performing well in classification.
- AUC (Area Under the Curve): Represents the area under the ROC curve (Receiver Operating Characteristic curve), which is a graph that shows the relationship between the True Positive Rate and the False Positive Rate across different classification thresholds. AUC measures the model's ability to distinguish between classes; the higher the AUC, the better the model can distinguish between positive and negative classes.

4.3 Results

Table 1. The results of model evaluation

Model	Accuracy	Precision	Recall	F1-score	AUC
PhoBERT + TF-IDF	0.888112	0.863014	0.913043	0.887324	0.922689
PhoBERT	0.872340	0.850649	0.909722	0.879195	0.947665
BERT	0.787234	0.850000	0.708333	0.772727	0.858343
RoBERTa	0.741135	0.844660	0.604167	0.704453	0.834541

Table 1 presents the evaluation results based on various metrics from the models when tested on the same training and testing datasets.

- Roberta: Roberta's classification performance is quite poor, with an Accuracy of 0.741 and an AUC of 0.835. While the Precision is 0.845, the Recall is only 0.604, showing that the model misses many instances of fake news. The F1 Score of 0.704 shows that although the model performs at a reasonable level, it is not as efficient as other models.
- BERT: Performs better than Roberta in classification, with an Accuracy of 0.787 and an AUC of 0.858. Precision is 0.850 and Recall is 0.708, showing a balanced performance between detecting fake news and legitimate news. The F1 Score of 0.773 shows that BERT is a strong model but still not the best among the tested models.
- PhoBERT: Achieves high performance with an Accuracy of 0.872 and an AUC of 0.948, showing excellent ability to distinguish between real and fake news. The F1 Score of 0.879 shows that this model has a good balance between Precision and Recall, although slightly lower than PhoBERT + TF-IDF.
- PhoBERT TF-IDF: Achieves the highest Accuracy of 0.888 among the models, with Precision (0.863), Recall (0.913), and AUC (0.923) all very good. This model balances accuracy and detection well, making it the most effective model for the task.

4.4 Discussion

The experimental results indicate that PhoBERT and PhoBERT + TF-IDF are both highly effective models for analyzing and classifying fake news. PhoBERT + TF-IDF achieved the best performance, with the highest Accuracy, Precision, and F1 scores, demonstrating its exceptional classification capabilities. PhoBERT also performed very well, with high Accuracy, Precision, Recall, and AUC scores.

However, it is important to note that while PhoBERT + TF-IDF has a slightly lower AUC compared to PhoBERT, it still maintains a high F1 Score, indicating a strong balance between Precision and Recall. This balance suggests that PhoBERT + TF-IDF may be more conservative, potentially missing some legitimate news but providing more accurate predictions overall.

Table 2. Comparison of the predictive performance of the models on the training dataset

	Posts/News		True label	Model's predicted label			
				RoBERTa	BERT	PhoBERT	PhoBERT + TF-IDF
1	"Tuyến Metro Nhổn Ga Hà Nội vận hành thương mại ngày 09/08/2024."	"Metro Nhon Ga Ha Noi line will operate on August 9, 2024."	Real	Real	Real	Real	Real
2	"Hà Nội gặp khó khăn di dời người dân ra khỏi vùng lũ."	"Hanoi faces difficulties in relocating residents from flood-prone areas."	Real	Fake	Fake	Real	Real
3	"Ban tổ chức Olympic hủy buổi tập 3 môn phối hợp lần thứ hai vì chất lượng nước sông Seine."	"The Olympic organizers canceled the second triathlon training session due to the water quality in the Seine River."	Real	Fake	Real	Real	Real
4	"Hiện trường kinh hoàng xe tải cố vượt đường ray khiến tàu hỏa trật bánh, ít nhất 100 người thương vong, hành khách hoảng loạn."	"A horrific scene unfolded as a truck attempted to cross the railway tracks, causing a train to derail. At least 100 people were injured or killed, and passengers were left in a state of panic."	Fake	Fake	Fake	Fake	Fake
5	"Tai nạn sập hầm lò đặc biệt nghiêm trọng ở Quảng Ninh khiến 5 công nhân tử vong,... đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân."	"A particularly severe mining accident in Quảng Ninh killed 5 workers,... the puppet party and trade unions have never cared about workers' conditions."	Fake	Real	Real	Real	Fake
6	"Sự cố nhánh cây dầu bị gãy, rơi từ độ cao khoảng 25m ở công viên Tao Đàn, TPHCM. Vụ việc làm 2 người tử vong, 3 người bị thương."	"An incident occurred in Tao Dan Park, HCM City, where a branch of an oil tree broke and fell from a height of about 25 meters. The accident resulted in 2 deaths and 3 injuries."	Real	Real	Real	Real	Fake

Table 2 presents some representative cases extracted from the training set. In most simple cases of real news and fake news, such as sentence 1, which is notification news, and sentence 4, which is fabricated information containing many words that attract attention, all 4 models give accurate results. However, for news cases with more information, models such as BERT and RoBERTa had many misclassifications, causing these two models to have low

performance and become unreliable.

Although PhoBERT and PhoBERT + TF-IDF demonstrated high accuracy in prediction, there were still some exceptions, particularly in cases where the news contained a mix of true and false information. For example, in case 5, “Tai nạn sập hầm lò đặc biệt nghiêm trọng ở Quảng Ninh khiến 5 công nhân tử vong,... đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân” contains the true information “Tai nạn sập hầm lò đặc biệt nghiêm trọng ở Quảng Ninh khiến 5 công nhân tử vong” (“A particularly severe mining accident in Quảng Ninh killed 5 workers”) but the additional part “…đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân” (“the puppet party and trade unions have never cared about workers' conditions”) is inaccurate and unverified. In this instance, PhoBERT + TF-IDF correctly classified it as fake news, while PhoBERT and other models were misled by the true part of the article. PhoBERT + TF-IDF's ability to accurately identify such cases is attributed to TF-IDF's emphasis on important keywords and its ability to minimize the influence of common but less informative words. TF-IDF helps the model recognize that additional information lacks validity and should not be considered real, thereby improving classification accuracy.

However, this cautious approach also led PhoBERT + TF-IDF to incorrectly classify some true news cases, such as case 6. For better results, it is necessary to effectively adjust the combination of PhoBERT and TF-IDF. This will help the model not only identify important keywords but also better understand the context surrounding those words.

5. Conclusions and Future work

In this study, we focused on leveraging Transformer models such as BERT, RoBERTa, and PhoBERT for fake news classification in Vietnam. We collected a dataset comprising Facebook posts from June to July 2024, covering topics such as lifestyle, society, and politics. Due to the limited number of fake news articles, we supplemented our dataset with additional fake news examples from the VFND dataset, as described in Ho Quang Thanh’s thesis, “VNFD - Vietnamese Fake News Datasets” [19]. We then applied Transformer models for classification, and the evaluation results showed that PhoBERT and PhoBERT combined with TF-IDF achieved the highest prediction performance for Vietnamese.

However, this model still has some limitations. One of the main problems is that the data is not enough and potential loss of information due to the way the Vietnamese language is structured, such as the use of acronyms, different grammar, or articles containing a mix of true and false information. This can lead to incorrect predictions from the model. Additionally, we currently only research and classify news based on the content of the news, without taking advantage of additional data such as the number of interactions and comments. These are important and quite large sources of information.

Moving forward, we plan to continue to collect data and incorporate analysis of comments on both true and false articles. This will help us better understand user sentiments and attitudes towards both types of information, ultimately contributing to more accurate prediction results.

References

[1] K. Chowdhary and K. R. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020. Article (CrossRef Link)

[2] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Adv Neural Inf Process Syst*, vol. 34, pp. 15908–15919, 2021. Article (CrossRef Link)

[3] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," *arXiv preprint arXiv:2003.00744*, 2020. Article (CrossRef Link)

[4] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. Article (CrossRef Link)

[5] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. Article (CrossRef Link)

[6] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. Article (CrossRef Link)

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. Article (CrossRef Link)

[8] A. Agarwal and P. Meel, "Stacked Bi-LSTM with attention and contextual BERT embeddings for fake news analysis," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2021, pp. 233–237. Article (CrossRef Link)

[9] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019. Article (CrossRef Link)

[10] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*, IEEE, 2019, pp. 518–527. Article (CrossRef Link)

[11] T. N. Hieu, H. C. N. Minh, H. T. Van, and B. V. Quoc, "ReINTEL Challenge 2020: Vietnamese Fake News Detection usingEnsemble Model with PhoBERT embeddings," in *Proceedings of the 7th international workshop on Vietnamese language and speech processing*, 2020, pp. 1–5.

[12] N.-D. Pham, T.-H. Le, T.-D. Do, T.-T. Vuong, T.-H. Vuong, and Q.-T. Ha, "Vietnamese fake news detection based on hybrid transfer learning model and TF-IDF," in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2021, pp. 1–6. Article (CrossRef Link)

[13] C.-V. Nguyen Thi, T.-T. Vuong, D.-T. Le, and Q.-T. Ha, "v3mfnd: A deep multi-domain multimodal fake news detection model for Vietnamese," in *Asian Conference on Intelligent Information and Database Systems*, Springer, 2022, pp. 608–620. Article (CrossRef Link)

[14] K. D. Pham, D. Van Thin, and N. L.-T. Nguyen, "Improving Vietnamese Fake News Detection based on Contextual Language Model and Handcrafted Features," *Science and Technology Development Journal*, vol. 26, no. 2, pp. 2705–2712, 2023. Article (CrossRef Link)

[15] T. O. Tran and P. Le Hong, "Improving sequence tagging for Vietnamese text using transformer-based neural models," in *Proceedings of the 34th Pacific Asia conference on language, information and computation*, 2020, pp. 13–20. Article (CrossRef Link)

[16] T. H. Vo, T. L. T. Phan, and K. C. Ninh, "Development of a fake news detection tool for Vietnamese based on deep learning techniques," *Eastern-European Journal of Enterprise Technologies*, vol. 119, no. 2, 2022. Article (CrossRef Link)

[17] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng*, vol. 69, pp. 1356–1364, 2014. Article (CrossRef Link)

[18] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indones. J. Electr. Eng. Comput. Sci*, vol. 22, no. 2, p. 976, 2021. Article (CrossRef Link)

[19] H. Q. Thanh and P. M. Ninh, "VFND: Vietnamese fake news datasets " GitHub, Feb. 2019. [Online]. Available: https://github.com/VFND/VFND-vietnamese-fake-news-datasets. [Accessed: Aug. 8, 2024].