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

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

Abstract

The spread of fake news on social media has become a serious issue, leading to misinformation and causing harm to society. **This research 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. **Nevertheless**, there are still challenges due to data imbalance between the number of true and false news. The posts were labeled as **real** or **fake**, then underwent data preprocessing and were trained using transformer models and PhoBERT for Vietnamese. We also incorporated the **TF-IDF and Word2Vec word embedding techniques** 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 paper outlines** the background, objectives, methodology, and future research directions, providing a comprehensive overview of the **research** 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, fake news is spreading quickly on social media, which has become a major societal issue, 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 acknowledged** as a potent instrument in the field of artificial intelligence in general, and in the field of natural language processing (NLP) in particular. 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 aims 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** increase the effectiveness and accuracy of identifying Vietnamese fake news. **Nonetheless**, 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 remainder of this paper is structured as follows: **In Chapter 2**, we briefly describe related works on transformer models and fake news detection, focusing on methods and models applied to research in Vietnam. Next, **in Chapter 3**, we present details of the proposed method, including the overall model and specific steps for system development. **Chapter 4** focuses on the experimental setup, dataset description, results, and discussion. Finally, **Chapter 5** provides a summary of our findings along with recommendations for further study.

2. Background

2.1 Related work

Detecting fake news is a topic that is being widely researched due to the rise 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. BERT (Bidirectional Encoder Representations from Transformers) was first presented by Devlin et al. in 2018 [5], this is 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 findings demonstrate that PhoBERT routinely outperforms the most recent top 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 (Term Frequency-Inverse Document Frequency) to generate word embeddings and uses CNN for feature extraction. This method achieved an AUC score of 0.9538. **That being said**, 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. **In spite of this**, 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 research** on detecting Vietnamese fake news on social media platforms using Transformer models, particularly PhoBERT and other BERT variants, it is necessary to master the following basic knowledge:

2.2.1 NLP

NLP is a branch of machine learning that enables computers to comprehend, analyze, and engage 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 task of text classification, NLP is utilized to extract information from the text, analyze its meaning, and convert the text into features that can be inputted into machine learning models or deep learning algorithms for the purpose of classification. For example, language processing techniques, such as bag of words, TF-IDF, and word embeddings, aid in the transformation of text into a digital format. 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 is composed of two primary components:

- 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

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

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 is trained on a diverse dataset of about 20 GB. This includes about 1 GB extracted from Vietnamese Wikipedia and the majority, about 19 GB, from Vietnamese news articles. 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

TF-IDF is a widely used method in Natural Language Processing (NLP) and text mining [17], [18]. It helps assess how significant a word is within 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 **method to supplement information about the frequency of word occurrences, which helps to enhance the ability to distinguish between important and unimportant words. This contributes to improving the accuracy of fake news classification without interrupting the context learning process of Transformer models.**

Specifically, TF-IDF is integrated as an additional layer into the features of the [CLS] token generated by Transformer models. This [CLS] token is the result of the hidden state from the final layer and represents the entire sentence, containing context information synthesized from all the words in the sentence. This approach will keep the process of learning context and the relationships between words in the sentence

completely handled by the Transformer without being affected by TF-IDF.

2.2.7 Word2Vec

Word2Vec is a technique that represents vocabulary as numerical vectors, typically built from large text datasets, and outputs a vector space that can have hundreds of dimensions. In this space, each word in the corpus is represented by a vector, and words with similar contexts and meanings, which tend to appear together, have vectors that are close to each other [19]**.**

We aim to combine Transformer models with Word2Vec to improve performance in tasks such as fake news classification by converting words into meaningful vectors. This helps the model understand and analyze words more accurately, reflecting semantic and contextual relationships. Integrating Word2Vec allows the model to gain deeper semantic information and improve the ability to process text data.

In this study, we use Word2Vec to initialize the Transformer's embedding layer, providing word vectors with semantic relationships as a starting point before fine-tuning on the training data. During the training process, this embedding layer will be adjusted to learn contextual features from the training data, while still maintaining semantic information from Word2Vec. This approach helps reduce computational costs, enhances the Transformer's ability to learn context, and optimizes performance for natural language processing tasks.

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 evaluate the model's accuracy, precision, recall, and F1 score using a distinct test dataset. Based on the evaluation results, we can further refine the model or adjust preprocessing techniques to enhance performance.

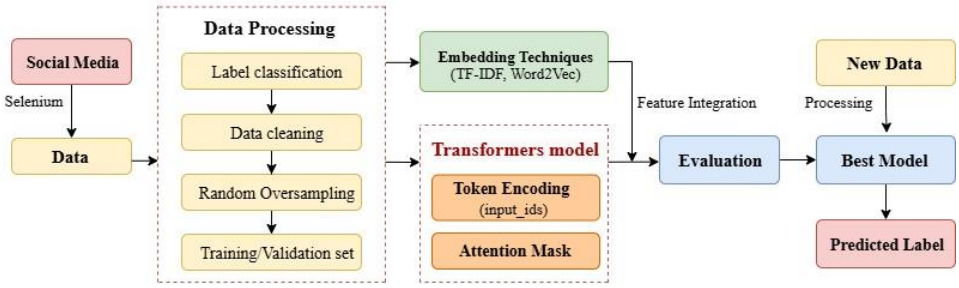

Fig. 1. General model of the system

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 illustrated in **Fig. 2**.

Despite this, 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 thú	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.fẽm quá" }, { "comment_id": "c20", "author": "Thang Vo",	
31/07/2024 10:50	https://www.facebook	Thương tâm quá: Trong lúc chở nhau trên xe máy v	0	https://www.fil": "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", "a	

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

3.3 Data processing

After the data collection step, we obtained 1,100 news items, including 812 that were labeled based on sources collected from reputable national sites or on information, knowledge, or events that have been proven to be true or false. The remaining 288 samples were manually labeled, inferred from available knowledge or community perspectives at the time the new was posted. Additionally, news related to superstitions or propaganda against the state and society were automatically classified as fake news during the labeling process. Finally, we collected 703 real news items and 397 fake news items for our dataset.

Then, 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" [20]. **Nonetheless**, 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 technique randomly copies existing samples of labels less than numbers until the number of samples of labels is balanced. **Even so, we did not copy the samples that were manually labeled by us.**

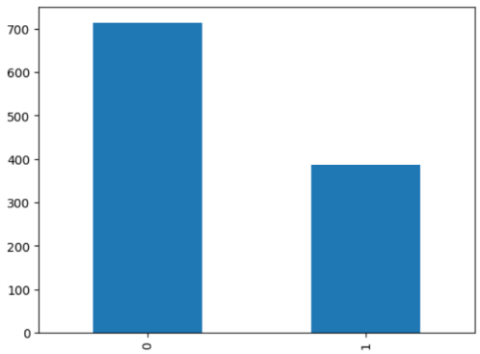

Fig. 4. Sample count for the two labels after collection

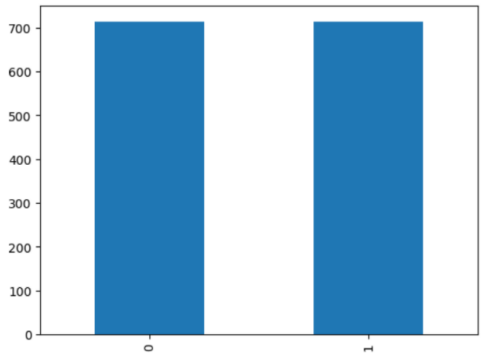

Fig. 5. Sample count for the two labels after processing

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 is real news and label 1 is fake news). Balancing the labels helps ensure the model does not become biased towards 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.

3.4 Transformer model training

In this phase, we train Transformer modelsafter obtaining the datasets. The three main models used are BERT, RoBERTa, and PhoBERT. For each model, we use the corresponding tokenizer to convert the text into numerical input sequences that the model can process.

We adjust the training techniques according to each model, including hyperparameter tuning and early stopping to optimize performance. During the training process, the models are monitored and evaluated regularly to ensure stable performance and avoid overfitting. We also experiment with different configurations to find the optimal setup for each model and achieve the most accurate results.

In addition, we also implement a method that combines PhoBERT with TF-IDF features. This method aims to maximize the semantic understanding ability of PhoBERT, while combining it with information about word frequency and importance through TF-IDF. We believe that this combination will significantly improve the ability to classify fake news compared to using PhoBERT alone.

Finally, we use the trained models to predict labels for the testing dataset and compare them with the actual labels to evaluate the performance of each model and compare them with each other.

3.5 An example

The models are trained using data from the training set, from which they learn the common characteristics of true and fake news. When encountering new news, the model will compare it with what it has learned to make an accurate prediction about whether the news is real or fake.

For example, training data includes news about diphtheria such as: “Hà Nội triển khai các biện pháp đề phòng bệnh bạch hầu xâm nhập” ("Hanoi implements measures to prevent diphtheria") and “Trong 6 tháng đầu năm 2024, Việt Nam ghi nhận 5 trường hợp mắc bệnh bạch hầu” ("In the first 6 months of 2024, Vietnam recorded 5 cases of diphtheria") which are real news, and “Tin sốc bệnh bạch hầu bùng phát với mức độ nguy hiểm chưa từng có, hàng trăm ca lây nhiễm mới ở Hà Nội.” ("Shocking news: Diphtheria outbreak with unprecedented danger level, hundreds of new infections in Hanoi.") which is fake new. The model learns to distinguish between the typical language patterns and sentence structures of real and fake news. In this case, the model recognizes that fake news often includes phrases emphasizing negative impact, such as “tin sốc” ("shocking news"), “sốc” ("shocking"), “hàng trăm ca lây nhiễm” ("hundreds of new infections"), “mức độ nguy hiểm chưa từng có” ("unprecedented danger level"), and so on, which are frequently found in fake news and are not common in real news articles.

When the model encounters new news, for example, “Sốc, hiện tượng chưa từng thấy. Hàng nghìn người xếp hàng chờ xét nghiệm bệnh bạch hầu” ("Shocking, unprecedented phenomenon. Thousands of people waiting in line for diphtheria testing"). Based on the information learned from the training set, the model can recognize that this new news has a similar structure and content to the fake news it has been trained on. The model pays attention to phrases like "sốc" ("shocking"), "hàng nghìn người" ("thousands of people"), “hiện tượng chưa từng thấy” ("unprecedented phenomenon") - terms that frequently appear in fake news. Therefore, the model is highly likely to classify this new news as "Fake."

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 **1,406 samples**, including both real and fake news across various domains. We then randomly split the dataset into an 80/20 training and testing set, resulting in 1,124 samples for training and 282 samples for testing. **In the training set, there are 232 manually assigned labels, and the testing set contains 56.**

This approach allows the model to learn patterns effectively, increasing the **probability** of better performance on new, unseen data. The dataset can be accessed at the following link: https://github.com/huynhtuan0106/Vietnamese-News-Dataset.

4.2 Evaluation tool

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 proportion of true positive samples out of all samples that the model has classified as positive. This reflects how accurately the model predicts positive cases.
- 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): This metric indicates the area under the ROC curve (Receiver Operating Characteristic curve), which charts the relationship between the True Positive Rate and the False Positive Rate at various classification thresholds. AUC evaluates how well the model can differentiate between classes; a higher AUC signifies a better ability to distinguish between positive and negative classes.

4.3 Results

Table 1. The results of model evaluation

Model	Accuracy	Precision	Recall	F1-score	AUC
RoBERTa	**0.741**	**0.845**	**0.604**	**0.704**	**0.835**
BERT	**0.787**	**0.850**	**0.708**	**0.773**	**0.858**
PhoBERT	**0.872**	**0.851**	**0.910**	**0.879**	**0.948**
PhoBERT + TF-IDF	**0.897**	**0.922**	**0.862**	**0.891**	**0.958**
PhoBERT + Word2Vec	**0.904**	**0.882**	**0.927**	**0.905**	**0.961**

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 good ability to distinguish between real and fake news. **Precision is 0.851, Recall is 0.910, and the F1 Score is 0.879, showing that PhoBERT performs better in Vietnamese news classification compared to RoBERTa and BERT.**
- PhoBERT + TF-IDF: **Achieves higher accuracy than PhoBERT, with an Accuracy of 0.897. The F1 Score (0.891) and AUC (0.958) are both good. The combination with TF-IDF helps PhoBERT reduce the misclassification of real news as fake, with the highest Precision of 0.922 among the models. However, Recall only reached 0.862, which is significantly lower than when using PhoBERT alone, leading to a higher risk as the model misses a considerable amount of fake news in the overall dataset.**
- PhoBERT + Word2Vec: **This model achieves the highest accuracy among all models, with an Accuracy up to 0.904, along with F1-Score (0.905) and AUC (0.961) — these are the three highest values compared to the other tested models, making it the most effective model for this task. Although its Precision is lower than PhoBERT + TF-IDF, its Recall reaches the highest value of 0.927. This shows that the model is more careful in classifying news, detecting more fake news overall, even though it misclassifies some real news as fake. This results in a lower risk for society.**

4.4 Discussion

The experimental results indicate that PhoBERT **and its combinations with TF-IDF and Word2Vec are highly effective models for analyzing and classifying fake news. PhoBERT + Word2Vec** achieved the best performance, with the highest Accuracy, Recall, and F1 scores, demonstrating its exceptional classification capabilities. **PhoBERT + TF-IDF** also performed very well, with high Accuracy and Precision.

It is important to note that while **PhoBERT + Word2Vec has slightly lower Precision than PhoBERT + TF-IDF, it achieves a higher Recall value. This suggests that PhoBERT + Word2Vec may be more conservative, potentially missing some legitimate news but offering more accurate overall predictions for fake news.**

Table 2 presents several cases extracted from the testting dataset that most models classified accurately, especially PhoBERT and its combinations with TF-IDF and Word2Vec. In most simple cases of true and fake news, such as the first news case, "Metro Nhon Ga Ha Noi line will operate on August 9, 2024", a straightforward true statement, and the third case, "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", a typical fake news example with fabricated content and attention-grabbing language, all models correctly classified these cases.

Table 2. Examples of correct classifications by most models in the testting set

	Posts/News	True label	Model's predicted label				
			RoBERTa	BERT	PhoBERT	PhoBERT + TF-IDF	PhoBERT + Word2Vec
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	Real
2	"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	Real
3	"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	Fake
4	**"Sốc, hiện tượng chưa từng thấy. Hàng nghìn người xếp hàng chờ xét nghiệm bệnh bạch hầu"** **"Shocking, unprecedented phenomenon. Thousands of people waiting in line for diphtheria testing"**	**Fake**	Real	Real	Fake	Fake	Fake

On the other hand, for more complex news, models like BERT and RoBERTa encountered numerous classification errors, leading to lower performance and unreliability. For instance, in case 2, "The Olympic organizers canceled the second triathlon training session due to the water quality in the Seine River", a true statement,

RoBERTa classified it as fake. Conversely, case 4, "Shocking, unprecedented phenomenon. Thousands of people waiting in line for diphtheria testing", which is fake, was classified as true by both RoBERTa and BERT.

Table 3. Examples of cases where the models misclassified

	Posts/News	True label	Model's predicted label				
			RoBERTa	BERT	PhoBERT	PhoBERT + TF-IDF	PhoBERT + Word2Vec
1	**"Nữ nhân viên làm việc tại Samsung lây nhiễm HIV cho 16 người"** **"A female employee at Samsung infected 16 people with HIV"**	**Fake**	Real	Real	Real	Fake	Real
2	"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 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	Fake
3	**"Trật bánh tàu hỏa tại ga Hải Vân Nam gây ùn tắc nhiều giờ liền, hàng trăm hành khách phải trung chuyển bằng xe khách"** **"A train derailment at Hai Van Nam station caused continuous hours-long traffic congestion, forcing hundreds of passengers to transfer by buses"**	**Real**	Fake	Fake	Fake	Fake	Real
4	**"Tin nóng tại Bình Thuận: Tai nạn xe khách nghiêm trọng làm tài xế tử vong, 11 người nhập viện cấp cứu"** **"Hot news from Binh Thuan: A serious passenger bus accident resulted in the driver's death and 11 people being hospitalized for emergency treatment"**	**Real**	Fake	Fake	Fake	Fake	Fake

Although **PhoBERT, PhoBERT + TF-IDF, and PhoBERT + Word2Vec** demonstrated high accuracy in prediction, there were still some exceptions, particularly in cases where the news contained a mix of true and false information**, as seen in the cases presented in Table 3.** Examples of cases where the models misclassified**. For example, in the first case: "A female employee at Samsung infected 16 people with HIV", a piece of information that has recently circulated widely on Vietnamese social media, only PhoBERT + TF-IDF correctly identified this as fake news, while the other four models mistakenly classified it as true.**

Another complex case of fake news is case 2: "A particularly severe mining accident in Quảng Ninh killed 5 workers,... the puppet party and trade unions have never cared about workers' conditions" contains the true information "A particularly severe mining accident in Quảng Ninh killed 5 workers" but the additional part "the puppet party and trade unions have never cared about workers' conditions" is inaccurate and unverified. In this instance, **PhoBERT + TF-IDF and PhoBERT + Word2Vec** correctly classified it as fake news, while PhoBERT and other models were misled by the true part of the article. **PhoBERT + TF-IDF showed its ability to focus on key terms and minimize the impact of common but less informative words. Meanwhile, Word2Vec helped PhoBERT better understand the full semantic meaning of the sentence, identifying the unverified, subjective addition and indicating that it should not be considered true, thereby improving classification accuracy.**

In case 3, "A train derailment at Hai Van Nam station caused continuous hours-long traffic congestion, forcing hundreds of passengers to transfer by buses", this is true information, but only PhoBERT + Word2Vec gave the correct result, the remaining models including PhoBERT + TF-IDF gave wrong prediction results. The reason could be that TF-IDF identified some keyword phrases that frequently appear in fake news from the training data, such as "hundreds of passengers" ("hàng trăm hành khách") and "continuous hours-long" ("nhiều giờ liền") which led to the incorrect classification. Relying on the frequency of words through TF-IDF sometimes affected the models' ability to understand the context, causing them to predict this as fake news, even though it was actually true.

In the final case, "Hot news from Binh Thuan: A serious passenger bus accident resulted in the driver's death and 11 people being hospitalized for emergency treatment", this is true information, but all models incorrectly predicted it as fake news. The cause might be that the news was presented with a tone and structure similar to other sensational fake news that the models had learned from before. Phrases like "Hot news" ("Tin nóng") and "serious" ("nghiêm trọng") carried an exaggerated tone and attracted attention, which distorted the models' classification ability. Although these words appear in many fake news stories, they can also appear in real news in emergency or truly serious situations.

These case show that while Transformer models like PhoBERT have great potential, they can still be misled by mainstream news that carries exaggerated meanings. To improve performance, it is necessary to adjust the models effectively, such as optimizing parameters and adjusting the weight of keywords. This fine-tuning will help the models better distinguish between true and fake news, even when they contain exaggerated contexts or information that mixes both true and false elements.

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" [20]. We then applied Transformer models for classification, and the evaluation results showed that **PhoBERT and PhoBERT combined with TF-IDF or Word2Vec achieved higher prediction performance compared to the other models in the Vietnamese context.**

Nevertheless, this model still has some limitations. One of the main challenges stems from the lack of data 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. Article (CrossRef Link)

[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] T. Mikolov, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, vol. 3781, 2013. Article (CrossRef Link)

[20] 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].