Utilizing Transformer Models to detect Vietnamese fake news on Social media flatforms

……..

Abstract

The proliferation of fake news on social media has become a serious issue, leading to misinformation and societal harm. This project aims to develop a system for detecting Vietnamese fake news using transformer models, particularly PhoBERT—a version of BERT optimized for Vietnamese. To address this issue, I collected a dataset consisting of Vietnamese posts from the social media platform Facebook and several official Vietnamese news articles, covering topics such as lifestyle, news, and politics. However, there were challenges due to the imbalance between the number of real and fake news. The posts were manually labeled as real or fake, then preprocessed and trained using transformer models and PhoBERT for Vietnamese, followed by evaluating their performance using metrics like accuracy, recall, and F1-score.

Our results indicate that PhoBERT outperforms other transformer models in detecting Vietnamese fake news, achieving high accuracy and reliability. This report outlines the context, objectives, methods, and future research directions, providing a comprehensive overview of the project and its contributions to the field of fake news detection.

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

1. Introduction

The rapid spread of fake news on social media has emerged as a serious societal issue, causing widespread misinformation and potential harm. Detecting fake news has various applications, such as ensuring the reliability of news sources, protecting public opinion, and helping to maintain social stability. While there is extensive research on fake news, it primarily focuses on English-language news.

In recent years, deep learning has been recognized as a powerful tool in artificial intelligence, particularly in natural language processing (NLP). However, traditional deep learning models often rely on sequential data processing, which can be limiting when handling complex language tasks.

Transformers, a novel architecture, have revolutionized NLP by utilizing attention mechanisms that allow for more effective context and relationship processing within text. This makes Transformers especially valuable for tasks such as fake news detection, where understanding nuanced language and context is crucial.

In this study, we focus on leveraging transformer models to detect fake news. Specifically, we use PhoBERT, an advanced variant of the BERT model designed specifically for the Vietnamese language. Our goal is to develop an effective system for identifying fake news on social media platforms, such as Facebook—the most widely used social media platform in Vietnam.

By harnessing the strengths of PhoBERT, we aim to improve the accuracy and effectiveness of Vietnamese fake news detection. However, we face significant challenges due to the lack of large-scale datasets containing both real and fake Vietnamese news. We have conducted data crawling from legitimate Facebook pages of official Vietnamese news outlets and fake data sources from impersonation pages, anti-establishment sources, and misinformation sites covering various fields from social life to politics. To achieve this, we utilized several tools, such as Selenium for data crawling, 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 applicable to our study in Vietnam. Next, in Part 3, we detail the proposed methodology, including the overall model and specific steps taken to develop the system. Part 4 covers the experimental setup, describes the dataset, results, and discussion. Finally, Part 5 concludes the paper by summarizing our findings and outlining future research directions.

2. Background

2.1 Related work

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

The journey of fake news detection has advanced significantly with the development of Transformer models. Vaswani et al. (2017) introduced the Transformer architecture, utilizing self-attention mechanisms to process sequential data efficiently, thus laying the foundation for modern NLP.

Since then, numerous studies have employed Transformer models for fake news detection in English. Devlin et al. (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), which uses bidirectional attention to better understand word context. Building on this foundation, Liu et al. (2019) enhanced BERT with RoBERTa, improving training efficiency and performance on NLP benchmarks. Additionally, Sanh et al. (2019) proposed DistilBERT, a smaller and faster version of BERT, suitable for real-time applications.

In the context of fake news detection, Agarwal et al. (2021) integrated a Bi-LSTM layer with attention into contextual embeddings for classifying trustworthy news in English. Monti et al. (2019) studied graph neural networks, employing a four-layer Graph CNN network combining user activity and article information to predict news. Meanwhile, Qi et al. (2019) emphasized the role of visual content, introducing a multidomain visual neural network that uses CNN and RNN models to analyze image features, aiding in distinguishing between fake and real news.

In the Vietnamese context, Dat Quoc Nguyen et al. (2020) developed PhoBERT, a Transformer-based model pre-trained on a large corpus of Vietnamese text, setting a significant benchmark for Vietnamese NLP tasks.

Their results indicate that PhoBERT consistently surpasses the recently leading pre-trained multilingual model XLM-R (Conneau et al., 2020), setting new benchmarks in several Vietnamese-specific NLP tasks such as part-of-speech tagging, dependency parsing, named-entity recognition, and natural language inference.

Recent studies have focused on utilizing PhoBERT and other deep learning techniques for Vietnamese fake news detection. For instance, Cao Nguyen Minh Hieu et al. proposed a tool during the ReINTEL 2020 Challenge that combined PhoBERT embeddings with temporal and community interaction metrics (shares, likes, comments). Their StackNet model achieved an AUC score of 0.9521, ranking first on the ReINTEL leaderboard.

Ngoc-Dong Pham et al. (2021) proposed a hybrid method combining PhoBERT with TF-IDF for word embeddings and CNN for feature extraction, achieving a notable AUC score of 0.9538, though the reliance on the ReINTEL dataset may limit diversity. Cam-Van Nguyen Thi et al. (2022) introduced v3MFND, a deep multimodal fake news detection model integrating text, images, and videos to enhance accuracy; however, the model's complexity may affect its real-time applicability. Earlier, Khoa Dang Pham et al. (2021) studied the vELECTRA model with handcrafted features, achieving an AUC score of 0.9575 on the ReINTEL dataset; nevertheless, the dependence on handcrafted features may limit adaptability. Vo Trung Hung et al. (2023) used CNN and RNN models to classify news into four groups, achieving an 85% accuracy rate. The dataset size may limit the generalizability of their results.

These studies highlight the effectiveness of Transformer models, particularly PhoBERT, in detecting Vietnamese fake news. They also underscore the importance of combining textual data with multimodal and meta-data to improve performance, while pointing out challenges related to dataset size, diversity, and computational complexity that need to be addressed in future research.

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 machine learning technology that enables computers to interpret, interact with, and understand human language. NLP encompasses various tasks such as syntactic parsing, semantic analysis, entity recognition, and text classification.

In the context of text classification, NLP extracts information from text, processes semantics, and represents the text in feature forms suitable for input into machine learning or deep learning models. Techniques like Bag of Words, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings help convert text into numerical forms. Subsequently, machine learning models like Naive Bayes, SVM (Support Vector Machine), and others can be trained to classify text into categories such as positive or negative sentiment, spam or non-spam, and real or fake news.

2.2.2 Transformer Model

The Transformer model represents a breakthrough in NLP, introduced by Vaswani et al. (2017) in the paper *"Attention Is All You Need."* The highlight of the Transformer lies in its self-attention architecture, which allows the model to learn relationships between words in a sentence without adhering to the sequential order used in previous models like RNN or LSTM.

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

- **Encoder:** The Encoder receives a sequence of words as input and represents them as semantic vectors. Each Encoder comprises multiple sequential layers, with two main components in each layer: the self-attention mechanism and a feed-forward neural network. The self-attention mechanism enables the model to learn the semantic relationships between relevant words in a sequence while ignoring unrelated words. The feed-forward neural network processes these attended vectors to produce deeper semantic representations.
- **Decoder:** The Decoder has a similar structure to the Encoder, using self-attention for the target input. Additionally, it employs cross-attention to connect with the Encoder's output. This enables the Decoder to generate semantic representations based on both the initial input sequence and the previously generated output sequence.

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 designed to understand word context in both directions (left-to-right and right-to-left) within a sentence. BERT is trained on two primary tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

Masked Language Modeling (MLM): In this task, some words in a sentence are replaced with the [MASK] token, and the model is required to predict the masked words based on the surrounding context.

Next Sentence Prediction (NSP): This task asks the model to predict whether a given sentence is the next sentence of a preceding sentence, thereby improving the model's ability to understand relationships between sentences.

BERT has achieved outstanding results in various NLP tasks such as text classification, entity recognition, and question answering.

2.2.4 RoBERTa (A Robustly Optimized BERT Pretraining Approach):

RoBERTa is a variant of BERT that has been optimized to improve performance by removing the Next Sentence Prediction (NSP) task and using a larger training dataset. RoBERTa applies the masked language modeling (MLM) approach with enhancements in training and data. RoBERTa has demonstrated superior performance in NLP tasks such as text classification and entity recognition, owing to optimized hyperparameters and training data.

2.2.5 PhoBERT

PhoBERT is a variant of BERT that has been trained entirely on Vietnamese text data, allowing the model to better capture the semantic and syntactic characteristics specific to this language.

PhoBERT incorporates improvements from RoBERTa, such as eliminating the Next Sentence Prediction (NSP) task and using only Masked Language Modeling (MLM), while being trained on a large-scale dataset.

This approach enables PhoBERT to perform more effectively compared to BERT or RoBERTa models trained on other languages.

3. Proposed methods

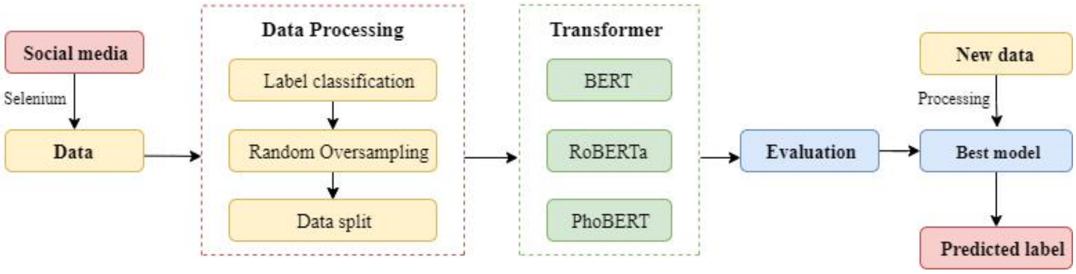
Our system can be divided into four main stages: (1) Data Collection, (2) Data Processing, (3) Model Training, and (4) Model Evaluation.

- **Data Collection:** In the first stage, we collect data from Facebook posts on official news pages as well as pages that spread fake, misleading, or disruptive content across topics like current affairs, lifestyle, and politics. We gather details such as the author, content, post link, and comments. This stage is crucial as the dataset will significantly influence the research outcomes.

- **Data Processing:** The collected data will undergo a series of preprocessing steps, including cleaning, text normalization, and the crucial step of manually labeling the posts as true or fake. After preprocessing, the data will be divided into training and testing sets and prepared for model training.

- **Model Training:** We use the processed data to train Transformer models: BERT, RoBERTa, and PhoBERT. We apply various training techniques to each model to optimize performance, including hyperparameter tuning and cross-validation techniques. After training, we compare the results of the three models to evaluate their effectiveness in detecting fake news.

- **Model Evaluation:** The final stage involves assessing 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 may further fine-tune the model or adjust preprocessing techniques to enhance performance.



4. Data Collection and Processing

4.1. Data Collection

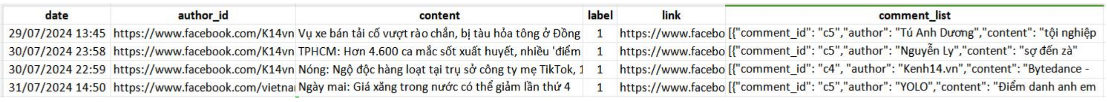
Due to the limitations of existing datasets on Vietnamese news and social media posts, and the outdated nature of available information, we decided to collect our own data to contribute to future research resources.

We manually selected and filtered posts. For authentic news, we identified reliable sources on Facebook, such as major Vietnamese news outlets and government pages. For fake news, we targeted sensationalist tabloids and Facebook groups that frequently share misleading information.

The data collection process faced many challenges, such as limited collection time and, in particular, difficulties in finding fake news sources, as some posts were deleted after being reported.

After selecting the necessary sources, we used Selenium to automate data collection, simulating user actions like navigating websites and extracting data.

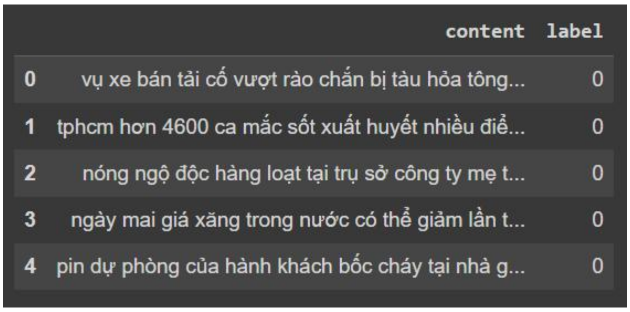
Ultimately, we collected two datasets: one for authentic news and one for fake news, with the latter being significantly smaller.



date	author_id	content	label	link	comment_list
29/07/2024 13:45	https://www.facebook.com/K14vn	Vụ xe bán tải cố vượt rào chắn, bị tàu hỏa tông ở Đồng	1	https://www.facebo	[{"comment_id": "c5","author": "Tú Anh Dương","content": "tội nghiệp
30/07/2024 23:58	https://www.facebook.com/K14vn	TPHCM: Hơn 4.600 ca mắc sốt xuất huyết, nhiều 'điểm	1	https://www.facebo	[{"comment_id": "c5","author": "Nguyễn Ly","content": "sợ đến zà"
30/07/2024 22:59	https://www.facebook.com/K14vn	Nóng: Ngộ độc hàng loạt tại trụ sở công ty mẹ TikTok, 1	1	https://www.facebo	[{"comment_id": "c4", "author": "Kenh14.vn","content": "Bytedance -
31/07/2024 14:50	https://www.facebook.com/vietnai	Ngày mai: Giá xăng trong nước có thể giảm lần thứ 4	1	https://www.facebo	[{"comment_id": "c5","author": "YOLO","content": "Điểm danh anh em

4.2 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 to be used.



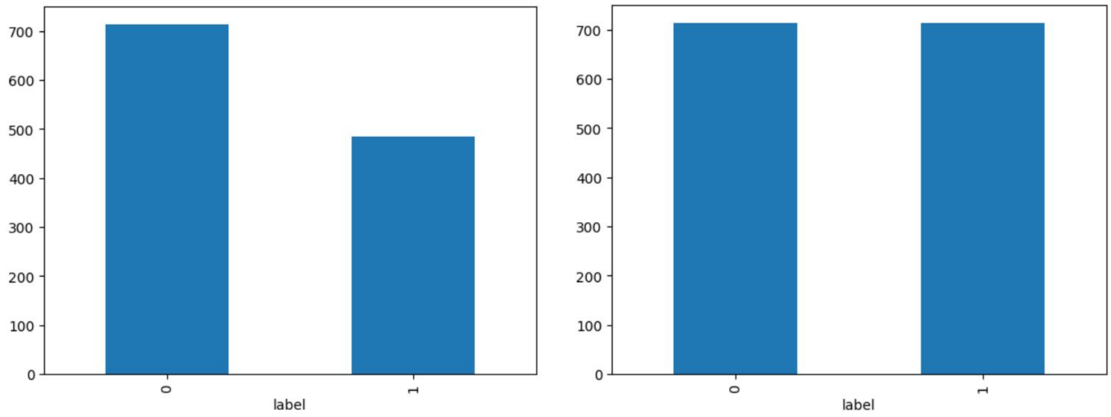
	content	label
0	vụ xe bán tải cố vượt rào chắn bị tàu hỏa tông...	0
1	tphcm hơn 4600 ca mắc sốt xuất huyết nhiều điể...	0
2	nóng ngộ độc hàng loạt tại trụ sở công ty mẹ t...	0
3	ngày mai giá xăng trong nước có thể giảm lần t...	0
4	pin dự phòng của hành khách bốc cháy tại nhà g...	0

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 evident that the number of fake news samples is significantly lower compared to true news, 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 integrated posts from the VFND dataset, as described in the thesis by Ho Quang Thanh, “VNFD Vietnamese Fake News Datasets: Tập hợp các bài báo tiếng Việt và các bài post Facebook phân loại 2 nhãn Thật & Giả.” However, since this dataset was collected in 2019, we selectively included only news that is not affected by time, such as scientifically debunked information, superstitious news, or distorted lifestyles. The supplemental data comprises no more than 20% of our current collection of fake news.

- Using Random Oversampling Technique: We applied the Random Oversampling technique from the ‘imbalanced-learn’ library. This method effectively balances the dataset by increasing the number of samples from the minority class. It works by randomly duplicating existing samples in the minority class until the class distribution is balanced.



After balancing the dataset, we divided it into two parts: a training set (1142 data rows) and a test set (286 data rows). We then encoded the data to prepare for training the Transformer models.

5. Experimental results

5.1 Evaluation tool

The classification results will be evaluated using Accuracy, Precision, Recall, F1 Score, and AUC.

- Accuracy: The ratio of correct predictions to total predictions, reflecting overall model performance, though it may not account for class imbalances.

- Precision: The ratio of true positive predictions to total positive predictions, indicating the correctness of positive predictions.

- Recall: The ratio of true positive predictions to actual positives, measuring the model’s ability to identify all positive instances.

- F1 Score: The harmonic mean of Precision and Recall, balancing these metrics, particularly useful for imbalanced data.

- AUC (Area Under the Curve): This represents the area under the ROC (Receiver Operating Characteristic) curve, 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; a higher AUC indicates better model performance in differentiating between positive and negative classes.

5.2 Results

Model	Accuracy	Precision	Recall	F1-Score	AUC
PhoBERT + TF-IDF	0.888112	0.863014	0.913043	0.887324	0.922689
PhoBERT	0.874126	0.835526	0.920290	0.875862	0.927977
BERT	0.758741	0.771654	0.710145	0.739623	0.837887
RoBERTa	0.522727	0.522727	1.000000	0.686567	0.399586

- Roberta: The classification performance is poor, with an Accuracy of only 0.5227 and a very low AUC of 0.3996. Although the model achieves a high Recall (1.0000), meaning it detects all legitimate news, this comes with a low Precision (0.5227), indicating that it incorrectly predicts many fake news items as real (false positives).

- BERT: Performs better than RoBERTa in classification, but still remains average, with metrics ranging between 0.7 and 0.8.

- PhoBERT: Achieves the highest performance, with an Accuracy of 0.874 and an AUC of 0.928, indicating excellent ability in distinguishing real from fake news. The F1 Score (0.876) shows that this model is well-balanced between Precision and Recall, though it is slightly lower compared to PhoBERT TF-IDF.

- PhoBERT TF-IDF: Achieves the highest Accuracy (0.888) among the four models, along with other metrics including Precision (0.863), Recall (0.913), and AUC (0.923), indicating that this model is well-balanced between accuracy and detection capability.

5.3 Discussion

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

However, it is important to note that while PhoBERT TF-IDF has a slightly lower Recall compared to PhoBERT, it still maintains a high F1 Score, reflecting a good balance between Precision and Recall. This balance suggests that PhoBERT TF-IDF is more cautious, possibly missing some legitimate news but providing more accurate overall predictions.

On the other hand, BERT did not perform as well as PhoBERT and PhoBERT TF-IDF. Although BERT showed fairly stable performance with average Accuracy, Precision, and F1 Score, the lower Recall and AUC indicate that it still has a higher rate of classification errors.

Roberta, on the other hand, is the weakest performer among the models. Although it has perfect Recall, meaning it detects all legitimate news, its very low Precision, Accuracy, and AUC scores indicate a large number of false positives, making Roberta's predictions of legitimate news unreliable.

6. 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: A Collection of Vietnamese News Articles and Facebook Posts Classified into Two Labels: Real & Fake." We then applied Transformer models for classification, and the evaluation results showed that PhoBERT provided the highest prediction performance for Vietnamese.

However, there are some limitations with this model, including the limited amount of data and some loss of information due to the structure of the Vietnamese language or the use of abbreviations and euphemisms in posts. This can lead to incorrect predictions by the models. Therefore, in the future, we will continue to collect data and incorporate additional analysis of comments on both real and fake posts to better understand user sentiments and attitudes towards the two types of information, ultimately contributing to more accurate prediction results.

References