

# A novel multimodal model for detecting Vietnamese toxic news using PhoBERT and Swin Transformer V2

Ngoc An Le<sup>1,2</sup>, Xuan Dau Hoang<sup>1</sup>, Xuan Hanh Vu<sup>2</sup>, Thi Thu Trang Ninh<sup>1</sup>

<sup>1</sup>Information Security Lab, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

<sup>2</sup>Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam

## Article Info

### Article history:

Received Feb 7, 2025

Revised Mar 26, 2025

Accepted Jul 2, 2025

### Keywords:

Detecting toxic news

Multimodal detection model

PhoBERT

Swin Transformer V2

Swin Transformer V2 +

PhoBERT

## ABSTRACT

News articles with fake, toxic or reactionary content are currently posted and spreaded very strongly due to the popularity of the Internet and especially the explosion of social networks and online services in cyberspace. Toxic news, especially reactionary news aimed at Vietnam, such as online articles spreading false information, slandering leaders, inciting destruction of the great national unity bloc, have a great impact on social life because they can spread quickly and have many forms of expression, such as news in the forms of text, images, videos, or a combination of text and images. Due to the seriousness of articles posting fake, toxic or reactionary news in cyberspace, there have been a number of studies in Vietnam and abroad for detection and prevention. However, most of the proposals focus on handling fake and toxic news posted using the English language. Furthermore, due to a large number of online news are posted in the form of images, or text embedded in images and videos, it is very difficult to process these news, leading to a relatively low detection rate. This paper proposes a multimodal model based on the combination of PhoBERT and Swin Transformer V2 for detecting fake and toxic news in both forms of text and images. Comprehensive experiments conducted on a dataset of 8,000 text and image news articles demonstrate that the proposed multimodal model surpasses both individual models and previous approaches, achieving 95% accuracy and 95% F1-score.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Xuan Dau Hoang

Information Security Lab, Posts and Telecommunications Institute of Technology

96A Tran Phu Rd., Ha Dong District, Hanoi, Vietnam

Email: dauhx@ptit.edu.vn

## 1. INTRODUCTION

The popularity of the Internet has made communication and information exchange more convenient and easier than ever. News and posts from users of social networks and online services can spread at breakneck speed in cyberspace. However, the rapid spread of user news and posts without adequate control brings great challenges, especially the problem of fake, toxic and reactionary news and posts. Specifically, these news and posts spread false information, slander political leaders, and even incite destruction of the great national unity bloc. These sabotage activities could lead to social instability and affect the national security. According to [1], more than 130,000 distorted articles and videos are spreaded on social networks and other platforms on the Internet each month on average in 2021, of which fake and toxic news and posts account for over 50%. Accordingly, there were more than 80,000 articles spreaded on Facebook, accounting for 67% and about 40,000 articles and videos posted on YouTube social network channels, personal blogs or other news sites. Taking advantage of the Internet and especially social networks, hostile forces have set up

thousands of news sites, blogs, hundreds of newspapers, publishers and radio and television stations with Vietnamese programs to distort and slander the communist party and the socialist regime in Vietnam. In this paper, we refer fake, toxic and reactionary news and posts as 'toxic news' for consistent references.

Research by Cao *et al.* [2] shows that posts with videos or images receive 18% more clicks, 89% more likes, and 150% more shares than posts without videos or images. To attract viewers, toxic news posted often have provocative, sensational content, stimulating viewers' curiosity. According to vtv.vn news site, on June 11, 2023, a serious terrorist attack occurred in Dak Lak province, Vietnam, killing 9 people, injuring 2 people, and burning down the government headquarters and many equipment of two communes. The underlying cause was that people believed in reactionary, anti-government news and posts from hostile forces abroad. This is a testament to the negative effects of news and posts with fake, toxic and reactionary content. Therefore, researching and detecting these toxic news on cyberspace, especially on social networks, is an urgent and practical task today.

Due to the seriousness of toxic news on cyberspace, there have been a number of studies on detection and prevention of this form of news, such as [3]-[13]. However, most studies focus on handling toxic news posted using the English language. Furthermore, due to a large number of news is being posted in the form of images, or text embedded in images, videos, or a combination of text and image content, it is highly difficult to process them, leading to a relatively low detection rate and a high false alarm rate. This paper proposes a multimodal model for detecting Vietnamese toxic news based on the combination of the PhoBERT model [14] and the Swin Transformer V2 model [15] to effectively handle two common forms of toxic news, including news in the form of text and images. The proposed model is able to distinguish toxic news from normal news better by using the PhoBERT model for recognizing text features and the Swin Transformer V2 model for recognizing image features in news. This paper's major contributions are:

- Proposing a model to detect toxic news in Vietnamese based on the combination of PhoBERT model and Swin Transformer V2 model;
- Collecting a dataset of Vietnamese toxic news, testing and evaluating the proposed model for detecting toxic news in Vietnamese.

The remainder of the paper is structured as follows: section 2 reviews related work in text and image news classification. Section 3 introduces various deep and transfer learning models, along with the proposed multimodal model. Section 4 provides the experimental scenarios, results and discussion. Finally, section 5 presents the paper's conclusion.

## 2. RELATED WORKS

This section introduces some closely related studies to the proposed model for detecting toxic news in the paper, including Nguyen and Gokhale [3], Uppada *et al.* [4], Armin *et al.* [5], Wu *et al.* [6] and Kiela *et al.* [16]. Among them, Nguyen and Gokhale [3] developed a model to identify anti-government sentiment on Twitter in English during politically driven anti-lockdown protests in Michigan's capital, USA. The model uses n-grams and term frequency-inverse document frequency (TF-IDF) techniques to extract and calculate values for features from news posts. The authors used algorithms such as random forests, support vector machine (SVM), logistic regression, distilled bidirectional encoder representations from transformers (DistilBERT) [17], multi-layer perceptron (MLP) to build the classifiers. The classifiers can effectively detect anti-government posts with an accuracy of about 85% and an F1-measure of about 82%. The drawback of this model is that it can only handle text posts and not image posts. In addition, the model was developed for processing English posts and is limited to the specific context of anti-lockdown protests in Michigan. This means that the model's classification performance may decrease when applying to posts or news in other languages or in other contexts.

In another approach, Uppada *et al.* [4] proposed a model that allows to identify articles with fake content. This model uses article data excerpt from Fakeddit English language dataset [7], which has over 1 million news samples containing metadata, text, image and title data collected from various sources. The paper uses a combined model of a component for text content processing and another component for image content processing in the articles. Specifically, BERT+Dense and RoBERTa+Dense are used for text content processing, while Xception [18], [19], Inception-ResNet-V2 [20], ResNet50 [21] and VGG19 [22] deep learning techniques are used for image content processing. The experimental results show that the model with the best detection performance is the combined model of Xception and BERT+Dense, with the accuracy of 91.94%, precision of 93.43%, recall of 93.07% and F1-score of 93%. The limitation of the model is that it is only trained on English language data, so the detection performance may decrease when applying to processing articles in other languages, especially Vietnamese.

Armin *et al.* [5] presented a multimodal model for detecting misinformation on social media using various data types of text, images, image comments and metadata in English language. The model's overall detection results are combined from results of individual detection on each data type using methods such as

sum, concatenate and maximum. Experiments confirm that the multimodal model produces 88% accuracy in the training phase and 88.1% in the testing phase using the combination of image and image comment data. In addition, the model achieves better detection results when using all types of data, including text, images, image comments and metadata of social media posts. However, similar to Uppada *et al.* [4], the proposed method is only trained on English language data, so the detection performance may decrease when applying to processing news and posts in other languages, such as Vietnamese.

Wu *et al.* [6] developed a multimodal fusion network model based on co-attention to detect fake news. This paper uses the BERT model for text processing and the visual geometry group-19 (VGG19) model for image processing. They use data collected from Twitter and Weibo, combining text features and image spatial and frequency features to detect fake news posts. The advantage of this model is its high detection performance, achieving 80.9% accuracy on the Twitter dataset and 89.9% on the Weibo dataset. However, the proposed model is only trained on English and Chinese language data, limiting its applicability to processing articles in other languages, such as Vietnamese. In addition, this model does not provide details on how to process each type of data and the co-attention mechanism, making it difficult to reproduce the model.

Kiela *et al.* [16] studied large-scale, high-speed multimodal model for news classification. The model is capable of handling multiple representations of news, such as discrete-mode with text, and continuous-mode with images derived from convolutional neural networks. In particular, the model focuses on scenarios that require rapid classification of large amounts of data. The authors also studied various methods for performing multimodal fusion and analyzed their trade-offs in terms of classification accuracy and computational efficiency. The results indicate that including continuous information improves performance over text alone on a variety of multimodal classification tasks, even with simple fusion methods. Research on multimodal news classification has opened up a promising direction for solving the problem of detecting fake and toxic news articles in cyberspace, because news articles are often posted in many expression forms, such as text, images, videos and using many different languages.

Thus, it can be seen that most of fake and toxic news detection models were developed for English, and therefore their direct application to Vietnamese news articles is fairly limited. Furthermore, the sufficient large datasets of Vietnamese news articles are not available, or not public. In addition, the detection accuracy of some proposals for toxic news detection is not high, such as of [3], [5], and [6]. This paper proposes a multimodal model aiming at improving the performance of detecting toxic news in Vietnamese based on the combination of PhoBERT and Swin Transformer V2 models.

### 3. PROPOSED MULTIMODAL MODEL FOR DETECTING TOXIC NEWS

#### 3.1. Overview of deep and transfer learning models

This section introduces some deep and transfer learning models used in previous and proposed multimodal models for toxic news detection. These deep learning multimodal models use a combination of text and image processing components in an integrated model. The text processing component is used to extract text features of news' text content while the image processing component is to extract image features of news' images. Then the text features and image features are combined to make the complete feature set for the next processing stages in the integrated model. Pre-trained models, such as BERT [23], [24], RoBERTa [25] and PhoBERT [14] are usually used in the text processing component while deep learning models, such as Xception [18], [19], VGG19 [22], Swin Transformer V2 [15] are used in the image processing component. Next part of this section will give a brief description of these deep and transfer learning models.

##### 3.1.1. BERT

BERT [23], [24] is a large-scale language model designed to comprehend and process text by capturing the context of each word within a sentence. Widely used in natural language processing (NLP), BERT excels in understanding word context, enabling more accurate text analysis. It delivers strong performance in various NLP tasks, including information extraction, text classification and machine translation. Trained on an extensive dataset, BERT can efficiently handle diverse types of text. However, as a large language model, it demands significant computational resources for both training and classification. Additionally, fine-tuning BERT for specific tasks can be complex and time-consuming.

##### 3.1.2. RoBERTa

RoBERTa [25] is an improved version of the BERT model, created by researchers at Facebook AI. Similar to BERT, RoBERTa is a transformer-based language model that leverages self-attention to analyses input sequences and generate contextualized word representations. A major distinction between BERT and RoBERTa is that RoBERTa was trained on a substantially larger dataset with a more refined training process.

Furthermore, it utilizes a dynamic masking strategy during training, enabling the model to develop more robust and generalizable word representations.

RoBERTa has proved exceptional performance compared to BERT and other cutting-edge models across various NLP tasks, such as text classification, question answering and language translation. Additionally, it has been used as a foundational model for various successful NLP models and is widely adopted in both research and industry applications. Similar to BERT, RoBERTa is a large language model, so it requires a high level of computational resources for training and classification. In addition, tuning RoBERTa for specific tasks can be complicated and time-consuming.

### 3.1.3. PhoBERT

PhoBERT [14] is a large-scale language model built on the BERT architecture and trained specifically on a Vietnamese dataset. PhoBERT is designed to understand and process Vietnamese text efficiently, helping to improve the accuracy of NLP applications for Vietnamese. The advantage of PhoBERT is that it is trained on a huge and diverse Vietnamese dataset, helping it understand the context and linguistic nuances of Vietnamese accurately. PhoBERT achieves robust performance across various Vietnamese NLP tasks, including text classification, information extraction, and machine translation. Similar to BERT, PhoBERT is a large language model, so it requires a high level of computational resources for training and classification. In addition, tuning PhoBERT for specific tasks can be complicated and time-consuming.

### 3.1.4. Xception

Xception [18], [19] is a convolutional neural network (CNN) architecture developed from the Inception V3 architecture, dedicated to image processing. It is an advanced model trained on a large image dataset, allowing for efficient recognition and classification of objects in images. Xception provides outstanding performance in image classification with high accuracy across a variety of datasets. It also has good generalization ability due to being trained on a large and diverse dataset. Xception is capable of accurately recognizing and classifying new objects that have never existed in the training dataset. Overall, the optimized architecture of Xception allows it to process images quickly and efficiently. The downside of Xception is that the architecture is complex, requiring a large amount of computational resources for the training and classification phases. In addition, Xception is also difficult to customize, therefore adjusting the Xception architecture to suit specific tasks can be complex and time-consuming.

### 3.1.5. Swin Transformer V2

Swin Transformer V2 [15] is an improved variant of the transformer model, which is specially designed for image processing. This model inherits the advantages of the previous version, while overcoming some limitations in performance and scalability. Swin Transformer V2 achieves better performance than the previous version in image classification, especially in processing high-resolution images. Swin Transformer V2 also has the ability to effectively process large-sized images, the ability to learn complex image features effectively, helping it achieve higher classification accuracy. The disadvantage of this model is that it requires more computational resources than traditional image processing models. In addition, optimization of Swin Transformer V2 to achieve optimal classification performance can be complex and time-consuming.

### 3.1.6. VGG19

VGG19 [22] is a deep CNN architecture with 19 layers, including 16 convolutional layers and 3 fully connected layers. The highlight of VGG19 is its simple structure, using small 3x3 filters, which helps extract features efficiently from low to high levels. Thanks to its uniform structure, VGG19 is easy to deploy and apply to many computer vision tasks, and achieves high performance on large datasets. However, VGG19 has the disadvantage of a large number of parameters (about 143 million), requiring a high level of computational resources, slowing down the training and inference process. The deep architecture is also prone to gradient vanishing, requiring special techniques to handle. Nevertheless, VGG19 is still an important pre-trained model, widely used as a foundation for many different applications thanks to its good feature learning ability and easy-to-understand structure.

## 3.2. The proposed model for detecting toxic news

### 3.2.1. Overview of the proposed model

The architecture of the proposed multimodal model for detecting toxic news is depicted in Figure 1. The proposed model uses a dataset of normal news and toxic news. Each type of news includes two forms: news presented in text form and news presented in image form. With image form, the text content of the news is integrated into the images. The proposed model is deployed in two phases: the training phase is shown in Figure 1(a) and the prediction phase in Figure 1(b). In the training phase, the processing flow of the

model is divided into two branches: the left branch uses the Swin Transformer V2 model to process news in image form and the right branch uses the PhoBERT model to process news in text form. The proposed model uses all layers of the fine-tuned PhoBERT and Swin Transformer V2 models, except for the final classification layers. The results of the two branches will be combined to produce a unified feature vector. The merged feature vector is fed into training to generate a prediction model. In the prediction phase, image and text news are pre-processed similarly during training to generate the merged feature vector and this vector is classified using the prediction model to produce the resulting label of the news as normal or toxic.

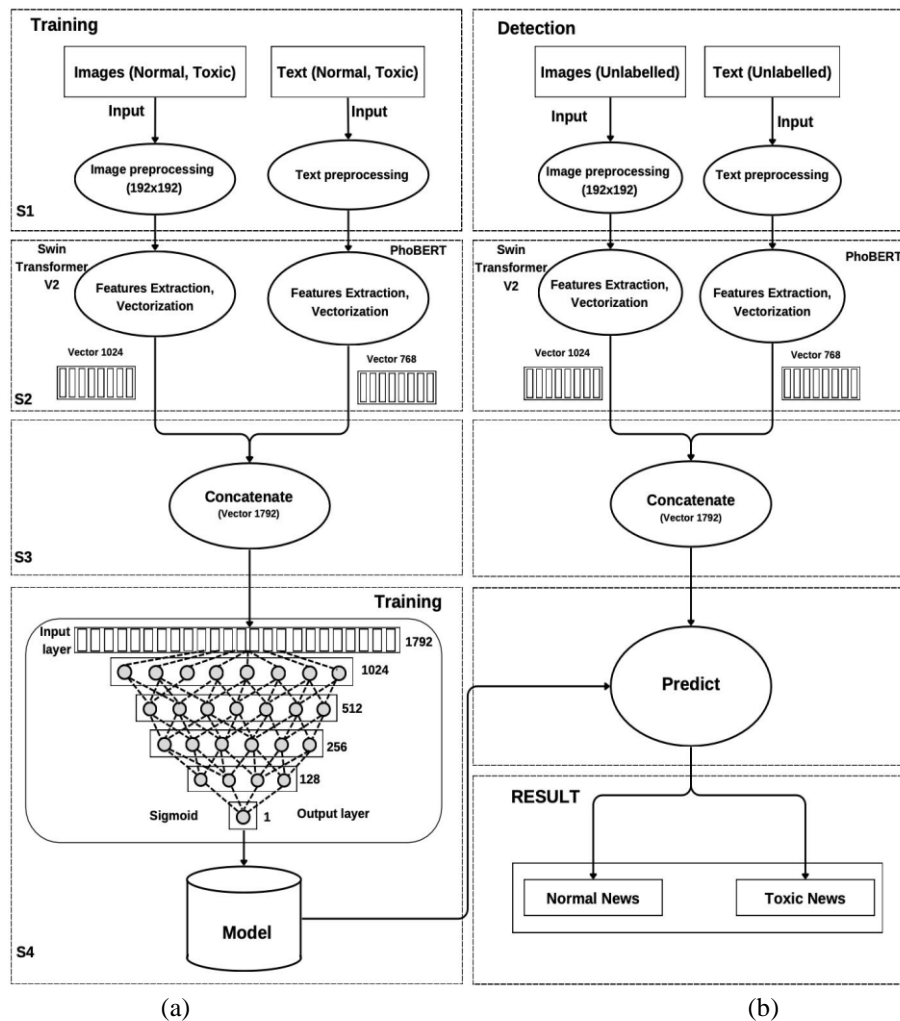


Figure 1. The architecture of the proposed multimodal model (a) the training phase and (b) the prediction phase

### 3.2.2. Data pre-processing

Data pre-processing is a crucial step in data processing, including both text-based and image-based news. With text-based news, especially when working with text automatically collected from the Internet, which often contains many unwanted characters and is in a non-uniform format. The first pre-processing steps include: text normalization, such as converting all letters to the same font, and removing strange characters. Next, remove unnecessary abbreviations, punctuation, and links. These characters often do not provide useful information for text analysis and can cause noise for machine learning models. The next important step in pre-processing is word segmentation, which is dividing the text into separate words to prepare for the next processing steps. The word segmentation process helps create suitable inputs for machine learning models, making it easier for them to analyze and understand the text content. Finally, the text data is padded to create fixed-length text strings that serve as input for training. For image news, the images are resized to 192x192 pixels before being fed into training.

### 3.2.3. Model training

The training process of the proposed multimodal model, as illustrated in Figure 1, consists of the following steps:

- Step 1 (S1): the news are preprocessed as described in section 3.2.2.
- Step 2 (S2): the preprocessed text news are fed into the PhoBERT model for processing and the outputs are vectors of 768 features. The preprocessed image news are fed into the Swin Transformer V2 model for processing and the outputs are vectors of 1,024 features.
- Step 3 (S3): combine each output vector of PhoBERT and each output vector of Swin Transformer V2 using the concatenate method, resulting in a synthetic vector of 1,792 features.
- Step 4 (S4): the 1792-feature synthetic vectors are used to train a classifier (Model) using a fully connected neural network with a hidden layer, 1 neuron in the classification layer with a sigmoid activation function. For model validation, the synthetic vectors are fed into the classifier or model to predict the label of normal or toxic news.

During the training process, the weights of the layers in both PhoBERT and Swin Transformer V2 branches are locked. This means that the weights learned in the previous training phase of these models will not change during the training of the multimodal model. The Adam optimization algorithm with a binary cross-entropy loss function is used and the training batch size is 32. Using the PhoBERT and Swin Transformer V2 models in the multimodal ensemble model offers many advantages over using each model individually. The multimodal model can effectively classify normal/toxic news thanks to its ability to combine the advantages of individual models, especially in complex data cases including news presented in both text and image formats.

### 3.2.4. Model prediction

In the prediction phase, the image and text news are preprocessed similarly during the training process in steps 1 to 3 to generate a combined 1792-feature vector. In step 4, the combined feature vector is classified using the prediction model (Model) to produce the result as a label of the news as normal or toxic.

### 3.2.5. Performance metrics

The proposed model's detection performance is measured using 4 metrics, including precision (PPV), recall (TPR), F1-score (F1), and accuracy (ACC). These metrics are computed using following formulas:

$$PPV = \frac{TP}{TP+FP} \quad (1)$$

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (3)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

where, TP, TN, FP and FN are parameters of the confusion matrix presented in Table 1.

Table 1. The confusion matrix and its TP, TN, FP and FN parameters

		Real class	
		Toxic	Normal
Predicted Class	Toxic	True Positives - TP	False Positives – FP
	Normal	False Negatives - FN	True Negatives - TN

## 4. EXPERIMENTS AND RESULTS

### 4.1. Collection of the experimental dataset

We have collected a dataset of news articles from social networks, forums and news sites for our experiments. The collected dataset consists of 8,000 Vietnamese news used to train and test the proposed multimodal model for detecting toxic news, including:

- News with toxic content include 2,000 articles in text form and 2,000 articles in image form collected from many different sources, such as on overseas forums, fanpages of social networks. News articles in this group are labelled 'Toxic';

- News with normal content consist of 2,000 articles in text form and 2,000 articles in image form collected from news sites and forums. News articles in this group are labelled ‘Normal’.

#### 4.2. Experimental results

To measure the performance of different combinations of text and image processing components in multimodal models, the following combination options have been selected:

- The proposed model: Swin Transformer V2 and PhoBERT
- Other multimodal models: Xception and PhoBERT, VGG19 and PhoBERT, Swin Transformer V2 and BERT, Xception and BERT, VGG19 and BERT, VGG19 and RoBERTa, Swin Transformer V2 and RoBERTa, Xception and RoBERTa.

The collected dataset is randomly split into three sections: 60% for training, 20% for validation, and 20% for testing, enabling performance evaluation of the proposed model alongside other multimodal models. Table 2 gives the detection performance of proposed model and other multimodal models using text and image features.

Table 2. Performance of the proposed model and other multimodal models

Multimodal models	Features	ACC (%)	PPV (%)	TPR (%)	F1 (%)
<b>Swin Transformer V2 and PhoBERT</b>	<b>Text+Image</b>	<b>95.0</b>	<b>95.0</b>	<b>95.0</b>	<b>95.0</b>
Xception and PhoBERT	Text+Image	93.0	93.0	93.5	93.0
VGG19 and PhoBERT	Text+Image	94.0	93.0	93.0	94.0
Swin Transformer V2 and BERT	Text+Image	90.0	85.0	89.5	89.5
Xception and BERT	Text+Image	86.0	85.5	86.0	86.0
VGG19 and BERT	Text+Image	88.0	87.5	87.0	87.0
VGG19 and RoBERTa	Text+Image	78.0	78.0	78.0	78.0
Swin Transformer V2 and RoBERTa	Text+Image	86.0	86.5	86.5	86.5
Xception and RoBERTa	Text+Image	81.0	81.5	82.0	81.5

#### 4.3. Discussion

From the experimental results given in Table 2, some comments can be made:

- The combined models using PhoBERT (Xception+PhoBERT, VGG19+PhoBERT) give significantly better detection performance than other combined models. This can also be seen that PhoBERT gives significantly better detection performance than BERT, RoBERTa because PhoBERT is pre-trained on a large Vietnamese dataset, therefore it is capable of processing Vietnamese news more effectively.
- Although the combined models using BERT/RoBERTa have lower detection performance than PhoBERT, the performance measurements are still pretty good. This is because news in the form of images often have more noise, containing less information than news in the form of text.
- In addition, it can also be seen that the combined models based on Swin Transformer V2 gives significantly better detection performance than those based on Xception and VGG19, because the Swin Transformer V2 model has the ability to process text data embedded in images better than the Xception and VGG19.
- The proposed multimodal model based on PhoBERT and Swin Transformer V2 uses text and image features produces superior detection performance compared to other multimodal models. Specifically, the performance metrics of the proposed combination model are accuracy (ACC) of 95%, precision (PPV) of 95%, recall (TPR) of 95%, and F1-score (F1) of 95%, which are much higher than the performance metrics of models based on other combinations.

Table 3 provides a comparison of the detection performance of the proposed multimodal model and related researches [4]. It is evident that the proposed multimodal model has significantly higher detection performance than the two proposed multimodal models in [4], including BERT+Xception (concatenate) and BERT+Xception (maximum). Specifically, the F1 measure of the proposed model based on PhoBERT+Swin Transformer V2 and the models based on BERT+Xception (concatenate) and BERT+Xception (maximum) [4] are 95% and 93.25% and 93.29%, respectively.

The detection performance of the proposed multimodal model is better than that of related research models because of the following two reasons: (i) the PhoBERT model has the ability to process Vietnamese data much better than the BERT, RoBERT models, and the Swin Transformer V2 model also has the ability to process text data embedded in images better than the Xception, VGG19 models. Therefore, the selection of PhoBERT and Swin Transformer V2 for the proposed combined model is appropriate, and (ii) choosing to combine the PhoBERT model with the Swin Transformer V2 model exploits the strengths of both individual models, helping to increase the ability to recognize features both on text and on image news, thereby being able to better distinguish between toxic news and normal news.

Table 3. Performance comparison of proposed model and previous models

Multimodal models	Features	ACC (%)	PPV (%)	TPR (%)	F1 (%)
BERT+Xception (concatenate) [4]	Text+Image	91.70	93.39	93.29	93.25
BERT+Xception (maximum) [4]	Text+Image	91.68	93.76	92.83	93.29
<b>PhoBERT+Swin Transformer V2</b>	<b>Text+Image</b>	<b>95.00</b>	<b>95.00</b>	<b>95.00</b>	<b>95.00</b>

## 5. CONCLUSION

This paper proposes a multimodal model based on the combination of the PhoBERT model and the Swin Transformer V2 model using features extracted from text news and image news. Experimental results on a Vietnamese news dataset, including 2,000 toxic news in text form, 2,000 toxic news in image form, 2,000 normal news in text form and 2,000 normal news in image form, confirm that the proposed combined model produces superior detection metrics compared to other combination models and existing multimodal models.

In the future, we will continue to research in the field of analysis and detection of toxic news, by adding new features into the proposed model in order to (i) continue to improve the detection accuracy and reduce the false alarm rate, and (ii) reduce the requirements of computational resources in training and especially in the detection of toxic news to improve the applicability in practice.

## ACKNOWLEDGEMENTS

The authors sincerely appreciate the invaluable support from the Information Security Lab at the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam and the Lab of Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam, in completing this project.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ngoc An Le	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Xuan Dau Hoang	✓	✓		✓	✓	✓			✓	✓		✓		
Xuan Hanh Vu			✓	✓		✓				✓	✓			
Thi Thu Trang Ninh		✓	✓			✓				✓	✓			

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [Xuan Dau Hoang], upon reasonable request.

## REFERENCES




- [1] T. D. Tran, "Electronic newspapers fight against false and reactionary arguments on social networks today [translate from the original language's name]: Bao dien tu dau tranh phan bac nhung luan dieu sai trai, phan dong tren mang xa hoi hien nay," 2021. <https://mst.gov.vn/bao-dien-tu-dau-tranh-phan-bac-nhung-luan-dieu-sai-trai-phan-dong-tren-mang-xa-hoi-hien-nay-197150305.html>.







- [2] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media*, Springer International Publishing, 2020, pp. 141–161.
- [3] H. Nguyen and S. Gokhale, "An efficient approach to identifying antigovernment sentiment on Twitter during Michigan protests," *PeerJ Computer Science*, vol. 8, p. e1127, Nov. 2022, doi: 10.7717/PEERJ-CS.1127.
- [4] S. K. Uppada, P. Patel, and B. Sivaselvan, "An image and text-based multimodal model for detecting fake news in OSN's," *Journal of Intelligent Information Systems*, vol. 61, no. 2, pp. 367–393, Nov. 2023, doi: 10.1007/s10844-022-00764-y.
- [5] K. Armin, S. Djordje, and Z. Matthias, "Multimodal detection of information disorder from social media," in *Proceedings - International Workshop on Content-Based Multimedia Indexing*, Jun. 2021, vol. 2021-June, pp. 1–4, doi: 10.1109/CBIMI50038.2021.9461898.
- [6] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569, doi: 10.18653/v1/2021.findings-acl.226.
- [7] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection," *arXiv*, 2020, <https://arxiv.org/pdf/1911.03854>.
- [8] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information Processing and Management*, vol. 60, no. 1, p. 103120, Jan. 2023, doi: 10.1016/j.ipm.2022.103120.
- [9] J. Ahmed and M. Ahmed, "Online news classification using machine learning techniques," *IJUM Engineering Journal*, vol. 22, no. 2, pp. 210–225, Jul. 2021, doi: 10.31436/ijumej.v22i2.1662.
- [10] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," in *W-NUT@EMNLP 2019 - 5th Workshop on Noisy User-Generated Text, Proceedings*, 2019, pp. 34–41, doi: 10.18653/v1/d19-5505.
- [11] K. S. Nugroho, A. Y. Sukmadewa, and N. Yudistira, "Large-scale news classification using BERT language model: spark NLP approach," in *ACM International Conference Proceeding Series*, Sep. 2021, pp. 240–246, doi: 10.1145/3479645.3479658.
- [12] A. Ali, S. Azman Mohd Noah, U. Kebangsaan Malaysia, and U. Bangi Lailatul Qadri Zakaria, "A BERT-based model: improving crime news documents classification through adopting pre-trained language models," pp. 0–16, Mar. 2023, doi: 10.21203/rs.3.rs-2582775/v1.
- [13] B. Juarto and Yulianto, "Indonesian news classification using IndoBERT," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 1, no. 2, 2023, [Online]. Available: <https://www.ijisae.org/index.php/IJISAE/article/view/2654>.
- [14] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 2020, pp. 1037–1042, doi: 10.18653/v1/2020.findings-emnlp.92.
- [15] Z. Liu et al., "Swin Transformer V2: scaling up capacity and resolution," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2022, vol. 2022-June, pp. 11999–12009, doi: 10.1109/CVPR52688.2022.01170.
- [16] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, vol. 32, no. 1, pp. 5198–5204, Apr. 2018, doi: 10.1609/aaai.v32i1.11945.
- [17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *NeurIPS*, 2019, [Online]. Available: <https://arxiv.org/abs/1910.01108v4>.
- [18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Jul. 2017, vol. 2017-January, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
- [19] G. Boesch, "Xception model: analyzing depthwise separable convolutions," 2021, [Online]. Available: <https://viso.ai/deep-learning/xception-model/>.
- [20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, vol. 31, no. 1, pp. 4278–4284, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.
- [21] Geeksforgeeks, "Residual networks (ResNet) – deep learning," Retrieved, 2025, [Online]. Available: <https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/>.
- [22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR 2015*, 2015, [Online]. Available: <https://arxiv.org/abs/1409.1556>.
- [23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2019, [Online]. Available: <https://arxiv.org/pdf/1810.04805>.
- [24] A. Tam, "A brief introduction to BERT," Retrieved Dec 20, 2023, [Online]. Available: <https://machinelearningmastery.com/a-brief-introduction-to-bert/>.
- [25] Y. Liu et al., "RoBERTa: a robustly optimized BERT pretraining approach," *arXiv*, 2019, [Online]. Available: <https://arxiv.org/abs/1907.11692>.

## BIOGRAPHIES OF AUTHORS







**Ngoc An Le**    received a master's degree in information systems at Posts and Telecommunications Institute of Technology in 2021. He is currently a lecturer at the Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam. He is also a Ph.D. student at Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His research interests include network security, cyberspace security, NLP, and computer vision. He can be contacted at email: AnLN.NCS2023@stu.ptit.edu.vn or anln@hou.edu.vn.







**Xuan Dau Hoang**     is an Associate Professor at the Faculty of Information Security, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. He received a Ph.D. degree in computer science from RMIT University, Melbourne, Australia in 2006. Associate Professor Hoang is the Dean of the Faculty of Information Security and Head of the Cyber Security Lab, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. His current research interests include attack and intrusion detection, malware detection, web security, and machine learning-based applications for information security. He can be contacted at email: dauhx@ptit.edu.vn.



**Xuan Hanh Vu**     is a senior lecturer at the Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam. He received a Ph.D. degree in information systems from Posts and Telecommunications Institute of Technology, Hanoi, Vietnam in 2022. His current research interests include anomaly detection, malware detection, system and software security, network security, and machine learning-based applications for information security. He can be contacted at email: hanhvx@hou.edu.vn.



**Thi Thu Trang Ninh**     received a master's degree in information systems at Posts and Telecommunications Institute of Technology in 2018. She is currently a lecturer at the Faculty of Information Security, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. Her research interests include attack and intrusion detection, malware detection, network security monitoring, web security. She can be contacted at email: trangnt2@ptit.edu.vn.