COMBINING TRANSFORMER, Bi-LSTM AND CNN FOR EFFICIENT VIETNAMESE FAKE NEWS DETECTION

Chi Khanh Ninh[1], Trung Hung Vo[2], Duy Khanh Ninh[3]*

[1]*The University of Danang – Vietnam - Korea University of Information and Communication Technology, Vietnam*
[2]*The University of Danang – University of Technology and Education, Vietnam*
[3]*The University of Danang – University of Science and Technology, Vietnam*

*Corresponding author: nkduy@dut.udn.vn

(Received: May 19, 2025; Revised: June 22, 2025; Accepted: June 23, 2025)

DOI: 10.31130/ud-jst.2025.23(9B).496

Abstract - Fake news spreading widely on digital platforms presents challenges where the rapidly evolving information landscape lacks effective detection. This paper introduces a hybrid deep learning model that integrates Transformer, Bidirectional Long Short-Term Memory (Bi-LSTM), and Convolutional Neural Network architectures to jointly capture both semantic and structural characteristics of Vietnamese text for the task of fake news detection. A high-quality dataset comprising 2,336 manually annotated Vietnamese news articles was developed, spanning three major domains: Politics, Healthcare, and Society. The proposed model was trained and evaluated on this dataset and benchmarked against two widely adopted baseline models: Gated Recurrent Unit and Bi-LSTM. Experimental results indicate that the proposed model achieves superior performance, attaining an overall accuracy of 95.3% and an F1-score of 0.951. These findings underscore the efficacy of combining multi-layered linguistic representations in enhancing Vietnamese fake news detection and contribute a valuable annotated resource for future studies in this domain.

Key words - Fake news detection; Transformer; Bi-LSTM; CNN; text classification; deep learning; natural language processing.

1. Introduction

In recent years, the proliferation of fake news on digital platforms has posed serious societal threats, particularly in the context of sensitive events such as pandemics, elections, or political-social conflicts. Fake news not only distorts public perception but can also jeopardize national security and social stability [1]. As a result, the development of automated systems for fake news detection, based on deep learning and natural language processing (NLP), has garnered increasing attention from the research community [2].

This study proposes a novel architecture combining Transformer, Bidirectional Long Short-Term Memory (Bi-LSTM), and Convolutional Neural Networks (CNN) to effectively extract both semantic and structural features from Vietnamese text. A specialized Vietnamese dataset was constructed, consisting of articles from Politics, Healthcare, and Society domains, manually labeled as either real or fake news to ensure evaluation accuracy. Through training and comparative analysis with Gated Recurrent Unit (GRU) and Bi-LSTM baselines, we demonstrate the superiority of the proposed model for Vietnamese fake news detection.

2. Related works

Fake news detection has been extensively studied from various perspectives. Traditional approaches often rely on hand-crafted features combined with machine learning algorithms such as Support Vector Machine, Naive Bayes, or Random Forest [3], [4]. However, these methods exhibit limitations when applied to texts with complex semantics or deceptive reasoning.

With the rise of deep learning, modern architectures such as LSTMs and GRUs have been successfully employed to learn semantic representations from textual data [5]. In particular, the Transformer architecture [6] and BERT [7] have demonstrated remarkable performance in capturing long-range contextual dependencies. Notable datasets and models such as FakeNewsNet [8], LIAR [9], and BERT-based methods for English have achieved promising results.

For Vietnamese fake news detection, however, public datasets and research efforts remain limited. Early studies employing Bi-LSTM, CNN, or attention-based models have struggled to effectively handle the linguistic nuances of Vietnamese news [10]. Another approach for fake news detection in Vietnamese utilized a Knowledge Graph (KG) in combination with Graph Convolutional Networks, reaching an 85% accuracy rate on a dataset constructed from Vietnamese online newspapers [13]. Duong et al. proposed a novel model for Vietnamese fact-checking that integrates a KG, Datalog, and KG-BERT (a deep learning model trained on this KG) to overcome resource scarcity and accurately classify information by extracting triples from complex sentences and enriching a large Vietnamese dataset, achieving 95% accuracy [16]. Pham et al. solved the Vietnamese fake news detection problem on social network sites by leveraging the PhoBERT pre-trained language model with TF-IDF for word embedding and CNN for feature extraction, which achieved an outstanding 0.9538 AUC score on raw data from the ReINTEL dataset [17]. A recent study introduced a hybrid approach for reliable Vietnamese fake news detection on social media, utilizing a pre-trained vELECTRA model combined with handcrafted features and achieving state-of-the-art performance with a 0.9575 AUC score on the ReINTEL dataset [18].

The previous studies show that the Vietnamese fake news detection is still a challenging problem. In that context, this paper contributes a novel hybrid model combining Transformer, Bi-LSTM, and CNN, along with a manually annotated Vietnamese dataset, to improve fake news detection performance.

2.1. Transformer

The Transformer architecture (Figure 1), introduced by Vaswani et al., has achieved remarkable success in NLP tasks [6]. Unlike traditional recurrent models such as LSTM or GRU, Transformer leverages self-attention mechanisms to model contextual relationships between words in a sequence. This self-attention mechanism, specifically multi-head attention, allows the model to weigh the importance of different words in the input sequence when processing each word. The architecture primarily consists of an encoder-decoder structure, where the encoder maps an input sequence of symbol representations to a sequence of continuous representations, and the decoder generates an output sequence one symbol at a time. Both the encoder and decoder are composed of stacked identical layers, each containing multi-head self-attention and a position-wise fully connected feed-forward network, with residual connections and layer normalization applied around each sub-layer. Positional encodings are also added to the input embeddings to inject information about the relative or absolute position of tokens in the sequence, as the self-attention mechanism itself is permutation-invariant.

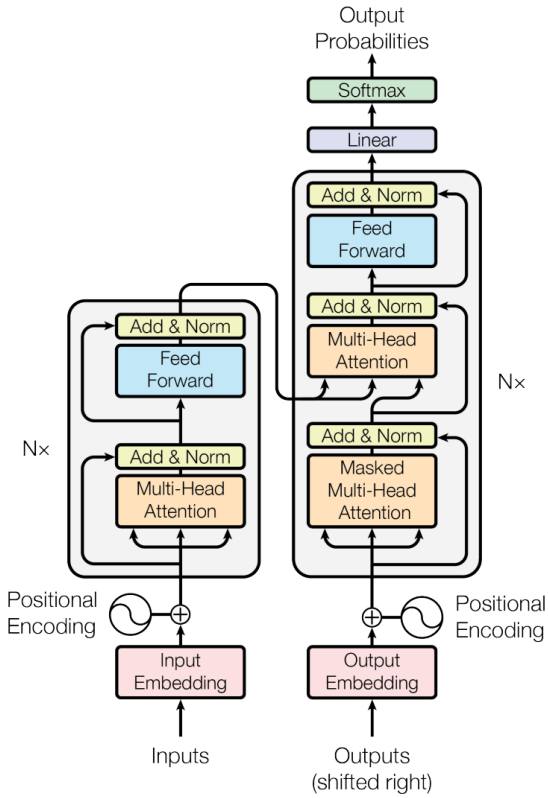

Figure 1. *The model architecture of Transformer [6]*

In the proposed model, only the encoder part of the Transformer is utilized, as it effectively captures dependencies among all tokens in the input text through self-attention and feed-forward layers. This enables the model to learn rich semantic and contextual representations of news documents. The decoder component is omitted, as the task does not require text generation. Position encoding is also incorporated to preserve word order, allowing the model to distinguish the relative positions of words without relying on sequential processing. By stacking multiple encoder layers, the model constructs comprehensive vector representations of the input, which serve as the foundation for subsequent feature extraction and classification.

2.2. Bi-LSTM

Bi-LSTM (Bidirectional Long Short-Term Memory) is a type of neural network designed for processing sequential data. It is a variant of the conventional LSTM architecture, engineered to capture contextual dependencies in both forward and backward directions - hence the term "bidirectional". The Bi-LSTM architecture comprises two parallel LSTM layers: one processes the input sequence in the forward direction, while the other processes it in reverse. This bidirectional structure enables the model to learn contextual information from both past and future input states, thereby facilitating the modeling of complex contextual dependencies [14].

In the proposed model, while the transformer layer enables the model to understand the semantics of individual words within a text, it may not fully capture the structural and contextual relationships throughout the sequence. To address this limitation, a Bi-LSTM layer is incorporated following the transformer module. This addition enhances the model's ability to learn and represent the overall context and structure of the input data more effectively.

2.3. CNN

Convolutional Neural Network (CNN) is a deep learning architecture widely used in image processing due to its ability to extract high-level features such as edges, contours, and textures. Inspired by the human visual system, CNNs consist of convolutional layers that apply filters across the input to detect salient patterns, followed by pooling layers that reduce dimensionality while preserving critical information. The extracted features are then processed by fully connected layers for tasks like classification or prediction [15].

In the proposed model, Convolutional Neural Networks (CNNs) are integrated to extract higher-level abstract features after the model has captured rich textual representations, encompassing both lexical and structural aspects through earlier layers. The use of CNNs in this context enables the model to analyze the text from multiple perspectives, facilitating the identification of latent or nuanced patterns that may be overlooked by purely sequential or context-based encoding mechanisms.

3. Proposed model

The proposed model is architected to effectively extract both structural and semantic features that differentiate fake news from authentic content. At the initial stage, a Transformer encoder is employed to generate token-level contextual representations via its self-attention mechanism, enabling the model to discern syntactic and semantic dependencies within sentences. To further capture the document-level structure and discourse flow, a Bi-LSTM network is integrated, allowing the model to learn sequential patterns and contextual transitions across sentences. Subsequently, a CNN layer is applied to extract higher-level

abstract features, analogous to how CNNs capture edges and textures in visual data, thereby revealing latent patterns often associated with deceptive content. The extracted feature representations are then passed through fully connected layers to perform the final binary classification, determining whether a given article is fake or real. The complete architecture of the proposed model is depicted in Figure 2.

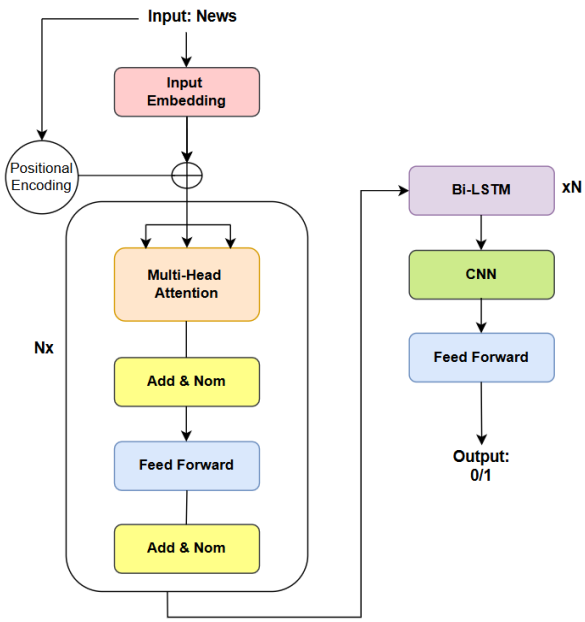

Figure 2. *Proposed model of fake news detection*

4. Result

4.1. Data collection

A dataset comprising over 2,000 Vietnamese news articles was curated to support the fake news detection task, spanning three primary domains: Politics, Healthcare, and Society. Articles identified as real were collected from well-established and credible Vietnamese news agencies, including *VnExpress*, *Dân Trí*, *Thanh Niên*, *VietnamNet*, and *Tuổi Trẻ*. Conversely, fake news samples were sourced from unverified or low-credibility platforms such as *quanlambao.blogspot.com*, *danlambaovn.blogspot.com*, *viettan.org*, *tinvn.info*, *nhanvanviet.com*, and *rfi.fr/vi*. Additional data points were incorporated from the publicly available VNFD dataset [11], which was developed specifically for Vietnamese fake news research.

Each article was manually annotated with binary labels: "0" for real news, verified through cross-referencing with trustworthy sources, and "1" for fake news, identified by the presence of misinformation or content distortion. The labeling process was conducted with strict quality control to ensure high annotation accuracy and maintain objectivity across the dataset.

4.2. Dataset description

The collected dataset comprises a total of 2,336 Vietnamese news articles, including 1,133 labeled as fake news and 1,203 as real news. The relatively balanced distribution between the two classes contributes to improved model performance by mitigating class imbalance during the training process. The distribution of real and fake news samples is presented in Table 1.

In terms of thematic categories, the dataset is divided as follows: 1,021 articles in the Politics domain, 1,083 articles in Healthcare, and 237 articles related to Social issues. Notably, political content accounts for the largest proportion due to its sensitive nature and broad national impact. Fake news in this category frequently originates from oppositional groups that utilize media as a tool to disseminate misinformation and destabilize political discourse.

In the Healthcare domain, the proliferation of fake news is largely driven by the context of the COVID-19 pandemic. Many fabricated articles were circulated to serve individual interests, such as promoting unverified medications, substandard medical equipment, or spreading false claims about government prevention policies. Articles categorized under Social issues tend to contain fabricated or misleading information intended to damage individuals' reputations or manipulate public opinion.

Table 1 and Table 2 summarize the annotated data distribution for the Vietnamese fake news detection task.

Table 1. *Distribution of Fake and Real news over three subsets*

Label	Training	Validation	Test
Real	724	241	241
Fake	679	226	226

Table 2. *Distribution of three domains over three subsets*

Domain	Training	Validation	Test
Politics	613	206	201
Healthcare	649	215	215
Society	141	46	46

4.3. Data preprocessing

Prior to model training, the collected dataset undergoes a series of preprocessing steps to standardize and clean the textual data, eliminate noise, and ensure consistency and suitability for the classification task. In this study, the preprocessing pipeline includes the removal of special characters, normalization of letters and digits, word segmentation, and stop word elimination. The stop word list utilized in this process is adopted from a widely used Vietnamese NLP resource [12]. Additionally, part-of-speech tagging is applied to support vocabulary construction and enhance the effectiveness of downstream model training.

4.4. Model training and evaluation

4.4.1. Hyper-parameter selection

Hyper-parameter selection is a critical phase, as it directly influences the model's performance. Choosing optimal hyper-parameters can significantly improve model effectiveness. For instance, by tuning parameters such as the learning rate, number of epochs, batch size, or model architecture, one can identify the best configuration to maximize performance. Additionally, appropriate hyper-parameter tuning helps prevent overfitting and under-fitting. Overfitting occurs when the model learns the training data too well and fails to generalize to unseen data, while under-fitting happens when the model fails to capture

the underlying patterns in the training data. Proper adjustment of hyper-parameters enables the development of a model with strong generalization capabilities.

In this study, key hyper-parameters selected for tuning include the number of attention heads, the number of encoder blocks in the Transformer, and the number of Bi-LSTM blocks. These components have a considerable impact on the model's overall performance.

To determine appropriate values for these parameters, empirical experiments were conducted. The experiments were trained using the binary cross entropy loss function and the Adam optimizer. Evaluation was based on the performance measured on the validation dataset. The parameter combinations were tested systematically: the number of attention heads was set to 1, 2, and 3; the number of encoder-Transformer blocks to 1 and 2; and the number of Bi-LSTM blocks to 1 and 2. All possible combinations of these values were explored (e.g., head = 1, encoder = 1, Bi-LSTM = 1; then head = 1, encoder = 1, Bi-LSTM = 2; etc.).

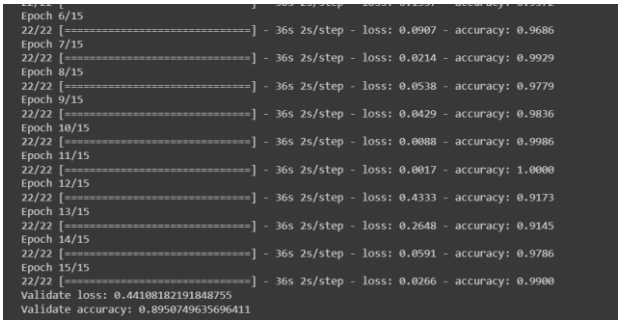

Figure 3. *The training process on Google's Colab*

Figure 3 shows the results of the training process on Google's Colab. The results indicate that the optimal configuration for the model consists of 3 attention heads, 2 encoder blocks, and 2 Bi-LSTM blocks.

4.4.2. Results

The proposed model was trained on Google's Colab with the following hyper-parameters presented in Table 3.

Table 3. *Hyper-parameters of model training*

Parameter	Value
Vocabulary size	28,841
Embedding dimension	16
Number of Transformer attention heads	3
Number of Transformer encoder blocks	2
Number of Bi-LSTM layers	2
CNN kernel size	3×3
Number of CNN layer	1
Activation functions	ReLU, Softmax
Loss function	Binary Cross-Entropy
Optimizer	Adam
Evaluation metrics	Accuracy, Recall, Confusion Matrix, F1-score
Training epochs	15

The training results are illustrated as follows. Figure 4 and Figure 5 illustrate the accuracy and loss curves on the training and the validation sets, respectively. It can be observed that the model started to be overfitted after the fifth epoch. Consequently, we selected the model at the fifth epoch as our optimal one to be used in the test phase. Table 4 depicts the confusion matrix of the proposed model on the test set.

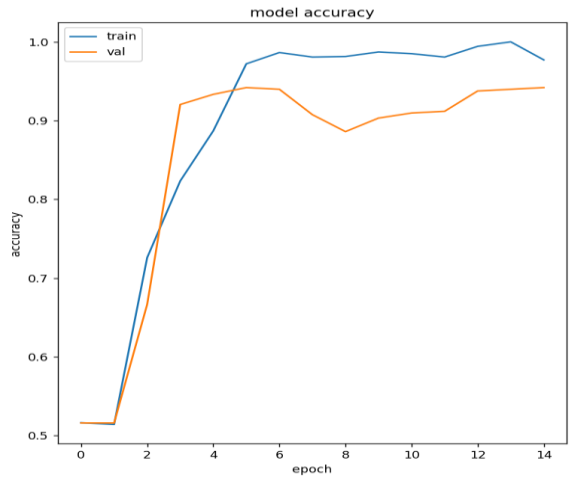

Figure 4. *Accuracy curve of proposed model*

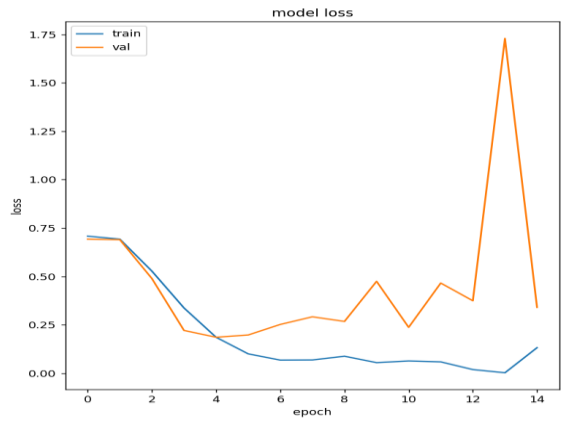

Figure 5. *Loss curve of proposed model*

Table 4. *Confusion matrix of proposed model*

	Predicted: Real	Predicted: Fake
Actual: Real	226	14
Actual: Fake	8	217

To objectively evaluate the effectiveness of the proposed model, we conducted a comparative analysis against two widely used deep learning architectures in NLP: GRU and Bi-LSTM. Both baseline models were trained on the same dataset using hyper-parameters configured to match those of the proposed model, thereby ensuring a fair and consistent evaluation. The comparison was performed on the test set, utilizing evaluation metrics such as accuracy, F1-score, and the confusion matrix. The confusion matrices corresponding to the GRU and Bi-LSTM models are presented in Table 5 and Table 6, respectively. A transformer-based architecture without Bi-LSTM was not included in this comparison since initial experiments have shown that the elimination of Bi-LSTM blocks from the proposed model make its detection performance on our dataset decrease.

Table 5. *Confusion matrix of GRU model*

	Predicted: Real	**Predicted: Fake**
Actual: Real	200	40
Actual: Fake	26	199

Table 6. *Confusion matrix of Bi-LSTM model*

	Predicted: Real	**Predicted: Fake**
Actual: Real	214	26
Actual: Fake	17	208

A comprehensive summary of the comparative results across all examined models is detailed in Table 7.

Table 7. *Model comparison results*

Model	**GRU**	**Bi-LSTM**	**Proposed**
Fake news accuracy (%)	88.4	92.4	96.4
Real news accuracy (%)	83.3	89.2	94.2
Accuracy score (%)	85.8	90.5	95.3
F1-score	0.857	0.905	0.951
Loss	0.95	0.77	0.24

5. Conclusion

In this paper, we proposed a hybrid model that integrates Transformer, Bi-LSTM, and CNN architectures for the task of Vietnamese fake news detection. The model is designed to jointly capture both semantic and structural features of text. By constructing a high-quality, manually labeled, and multi-domain dataset, we conducted a comprehensive training and evaluation of the model.

Experimental results demonstrate that the proposed architecture achieves an accuracy of 95.3% and an F1-score of 0.951, outperforming baseline models such as GRU and Bi-LSTM. These findings validate the effectiveness of combining multiple layers of linguistic representation in detecting fake news. For future work, possible extensions include incorporating social media signals, article metadata, and experimenting with large-scale pre-trained language models such as PhoBERT to further enhance system performance and generalization capability.

REFERENCES

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017. doi: 10.1257/jep.31.2.211

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective", *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36. https://doi.org/10.1145/3137597.313760

[3] S. I. Manzoor, J. Singla and Nikita, "Fake News Detection Using Machine Learning Approaches: A Systematic Review," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 230–234.

[4] V. Sivasangari, V. A. Pandian, and R. Santhya, "A modern approach to identify the fake news using machine learning", *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 3787–3795, 2018.

[5] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection", in *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM)*, New York, NY, USA, 2017, pp. 797–806.

[6] A. Vaswani et al., "Attention Is All You Need", in *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, CA, USA, 2017, pp. 6000–6010.

[7] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media*", Big Data*, vol. 8, no. 3, pp. 171–188, 2020. doi: 10.1089/big.2020.0062

[9] W. Y. Wang, ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* Vancouver, Canada, 2017, pp. 422–426.

[10] D. V. Vo and P. Do, "Detecting Vietnamese fake news", *CTU Journal of Innovation and Sustainable Development, vol. 15(Special issue on ISDS),* pp. 39–46, 2023.

[11] Q. T. Ho and M. N. Pham, "VFND - Vietnamese fake news datasets", *github.com*, February, 2019. [Online]. Available: https:// https://github.com/WhySchools/VFND-vietnamese-fake-news-datasets [Accessed March 03, 2025].

[12] V. -D. Le, "Vietnamese stopwords", *github.com*, 2015. [Online]. Available: https://github.com/stopwords/vietnamese-stopwords [Accessed March 13, 2025].

[13] V. D. K. Nguyen and P. Do, "Fake news detection using knowledge graph and graph convolutional network", *Journal of Intelligent and Fuzzy Systems*, vol. 45, no. 6, pp. 11107–11119, 2023. doi:10.3233/JIFS-233260

[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. doi: 10.1109/78.650093

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in *Proceedings of Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012, pp. 1097–1105.

[16] H. T. Duong, V. H. Ho, and P. Do, "Fact-checking Vietnamese Information Using Knowledge Graph, Datalog, and KG-BERT", *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 10, 240, 2023. https://doi.org/10.1145/3624557

[17] 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 *Proceedings of the 13th International Conference on Knowledge and Systems Engineering (KSE)*, Bangkok, Thailand, 2021, pp. 1–6.

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