Fake News Detection System, based on CBOW and BERT

Trung Hung Vo[1], Imre Felde[2] and Khanh Chi Ninh[3]

[1] The University of Danang – University of Technology and Education, 48 Cao Thang, Hai Chau, Da Nang, Vietnam, vthung@ute.udn.vn

[2] Óbudai Egyetem, Bécsi út 96/B, H-1034 Budapest, Hungary, felde.imre@uni-obuda.hu

[3] The University of Danang – Vietnam Korea University of Information and Communication Technology, Danang, Vietnam, nkchi@vku.udn.vn

Abstract: Fake news is becoming a major challenge that greatly affects the public's trust in the media. In this paper, we propose a new solution, combining word embedding based on CBOW (Continuous Bag Of Words) and the BERT (Bidirectional Encoder Representations from Transformers) models to support fake news detection. This paper focuses on presenting the proposed model and processing steps through the FND4Vn system, with a data set of Vietnamese news. Experimental results show that this solution achieves accuracy as high as 0.96 in recall and has many advantages compared to existing methods.

Keywords: Fake News Detection; CBOW; BERT; Transformer; Natural Language Processing

1 Introduction

Fake news is not a new phenomenon and has historically been commonly used in certain situations. The purpose of fake news is to defame or elevate individuals, solicit support for organizations or states and/or portray minority groups in an intentionally negative manner. The tactic of creating public opinion by providing false information has been used as a propaganda tool and has now become a major problem. Fake news is often related to politics, marketing goods, image promotion, elections and is especially popular in the period leading up to the 2016 US presidential election or the current Russia-Ukraine war.

Currently, readers tend to doubt the reliability of information, especially information posted on social networks. [1] shows that "using either mainstream or alternative news sources is associated with higher levels of trust in comparison with using social media as a main source". However, along with the strong development

of the Internet, users easily access a huge amount of information through many different communication channels, especially through social networks. This environment also brings about an explosion of fake news that makes it difficult for users to distinguish what is real and what is false. Therefore, more and more users lose confidence in the information received. Statistics for February 2023 of Statista (https://www.statista.com/) show the trust of users in the media:

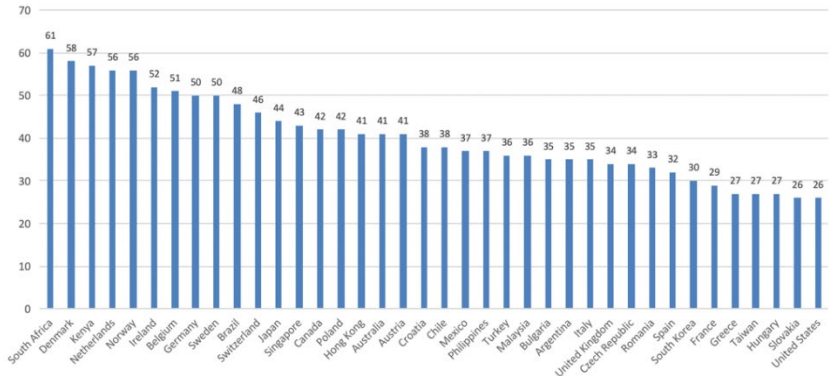

Figure 1
Percentage of respondents trusting news media (https://www.statista.com/)

In the past few years, there has been a lot of research related to detecting fake news and certain results have been achieved [2]. However, detecting fake news is a difficult problem because those who create and distribute fake news always try to adapt to create fake news that is most similar to real news and difficult to detect. Furthermore, whether a news story is determined to be fake or not depends largely on the perspective and perception of the evaluator. Therefore, research on detecting fake news continues to attract scientists. Key issues related to detecting fake news include building and updating data warehouses; how to process and represent data; techniques to improve correct detection rates and processing times.

This paper presents our research results, when building the FND4Vn system, to support the detection of fake news for two fields politics and pandemic on news written in Vietnamese. The new contribution of this study, is the proposal of a general model and detailed experimentation steps, based on the combination of CBOW and BERT. With CBOW, document is vectorized based on, not only the words appearing in that text, but also neighboring words (taking into account contextual factors). With the vector representation based only on frequency, this method allows increased accuracy because contextual factors in the text are also very important. With BEST, this technique also allows efficient calculations on multidimensional vector input data. The combination of these two techniques contribute to improving the effectiveness of content-based, fake news detection. The content of this paper includes main parts such as related research, proposal approach, experimental process, analysis and evaluation of experimental results, and finally, conclusions.

2 Related Research

Currently, there are many different approaches to building systems to support fake news detection. These approaches include analysis of news content, analysis of news origin, analysis of author's style, analysis of news distribution route, etc. [3]. Among them, the content-based analysis method is the most commonly used. One of the most commonly used content-based fake news detection methods today is deep learning. [4]. If we search with Google for the term *"fake news detection"* and *"deep learning"*, we get the results… *"About 156,000 results (0.37 seconds)"*.

Deep feedforward neural networks are the basic model of deep learning. The goal of a feedforward network is to approximate some function f^*. For example, for a classifier, $y=f^*(x)$ maps input x (input) to output y (output). The forward neural network defines a mapping $y=f(x,b)$ and finds the values of the parameters b, resulting in the best approximation of the function f.

The general model of the neural network is shown below:

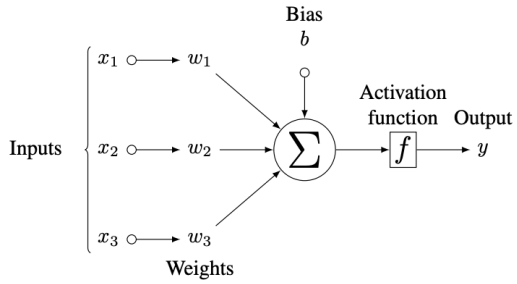

Figure 2
Example of neural network model

The basic model of a neuron is called a receptor. Perceptron receives input signal $x= (x_1, x_2, ..., x_{n+1})$ through transition layers to generate vector $w=(w_1, w_2, ..., w_{n+1})$. The Perceptron output is given as the dot product of the weight and the vector, transformed by the activation function:

$$output = f(w.x) = f(\sum_{i=1}^{n+1} w_i x_i) \quad (1)$$

Based on this general model, one can propose various deep learning algorithms that behave similarly to machine learning algorithms. However, there is a key difference that deep learning algorithms have different layers of data interpretation. Artificial Neural Networks refers to networks of such algorithms collectively known as Perceptrons [5].

Depending on the data, purpose and available resources, one can choose to use different deep learning solutions such as Convolutional Neural Network (CNN) [6], Recurrent Neural Network (RNN) [7], Gated Recurrent Unit (GRU) [8], Long Short-Term Memory (LSTM) [9],...

The advantage of this method is that it effectively exploits the storage and computational capacity of computer systems to analyze the content of documents and extract features of the data to predict an information is fake or real. However, this approach also has the limitation that fake news, real news arises very large in real time, this requires training data to create predictive service models that must also be constantly updated to ensure accuracy.

3 Proposal

3.1 General Model

The FND4Vn (Fake New Detection for Vietnamese) includes two components corresponding to two phases: training and detection. The training phase includes data collection, data preprocessing, vectorization based on Word Embeding, and training to create the model. The detection phase includes preprocessing the text to be evaluated, vectorization, and evaluation based on the model created in the previous stage. The general architecture of FND4Vn is as follows:

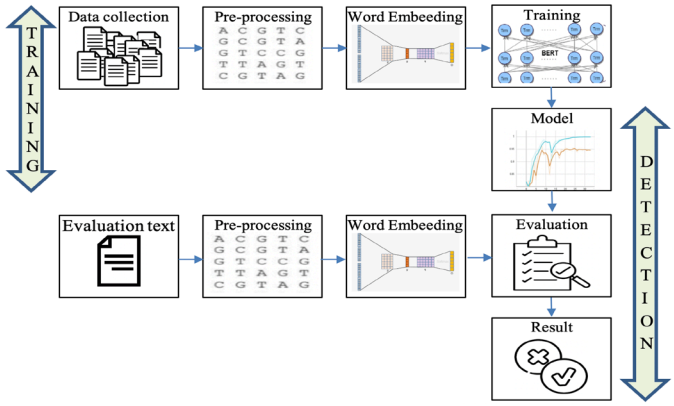

Figure 3
General architecture of the fake news detection system

Data collection: Data collection (including real and fake news in Vietnamese) is an important step in the process of building a classification model. Data is collected from online news sources. After collection, the data is selected and labeled by experts.

Pre-processing: In this step, it is necessary to remove special characters, standardize content, separate words, remove stop words, and perform word classification to build a vocabulary.

Word Embedding: To represent text into vector space based on Continuous Bag of Words (CBOW) technique, using neural networks to train context data.

Training: To build classification models. Within the scope of this research, the BERT model is focused on studying and comparing results with regression models applied in other fake news classification problems such as LSTM, GRU, BiLSTM and CNN. Each model has different advantages and disadvantages, requiring testing and comparison to choose the appropriate model. After building the model, the next step is to train and evaluate the effectiveness of the model on the training dataset and test dataset. Use metrics such as accuracy, recall, precision, F1-score, and ROC-AUC to evaluate model performance. Based on the evaluation results, steps will be taken to optimize the model by adjusting parameters, model structure, or data preprocessing techniques. This process requires testing and comparison between different models to find the best model for classifying fake news in Vietnamese.

Evaluation: After optimizing and selecting the appropriate model, the next step is to conduct testing for practical application. This experimental tool will help users distinguish fake news from accurate news based on entering the title and text content of a news to be evaluated.

3.2 CBOW

The model allows predicting the target word based on the context of surrounding words [10] [11]. CBOW model architecture as presented here:

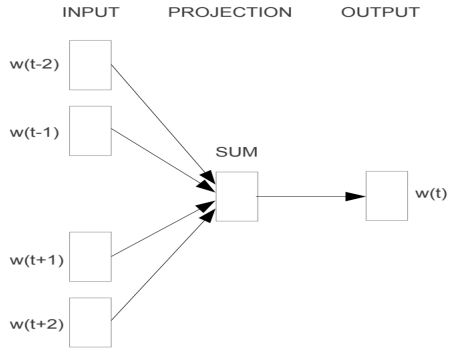

Figure 4
CBOW model architecture

This is a model that converts a sentence into word pairs of the form *(context word, target word)*. Depending on the purpose, we will have to set the window size. For example, for the sentence *"This is the context model"* and the window for the context word is 2 then the word pairs will be *([This, the], is)*, *([is, context], the)*,*([the, model], context)*. With these word pairs, the model will try to predict the target word as the context word.

The main idea of this model is that neighboring words appearing close to each other in a text will have very similar meanings again and again. Therefore, a central word is expected to appear conditional on the occurrence of neighboring words. The probability p is expected as the formula below:

$$p = p(c|w_1)p(c|w_2) \dots p(c|w_n) \quad (2)$$

In this formula: c is the centre word and w_i are words that surround the centre word.

The model can be proved to perform well as a result of a big probability p. The goal is to achieve the maximum objective function by maximizing:

$$obj = argmax \sum_{w \in text} \sum_{c \in context(w)} p(c|w;\theta) \quad (3)$$

in formula (2), θ is a hyper-parameter.

CBOW can calculate the similarity between word pairs and relatedness tasks. Therefore, it is appropriate to use CBOW in the content analysis process to detect fake news. In the past, people often used one-hot encoding to represent words. However, when the size of the corpus is too large, it will encounter serious problems and take a lot of time to train. With the proposal of CBOW, most of the problems related to that are now fixed.

3.3 BERT

Bidirectional Encoder Representation from Transformers (BERT) was developed by Google and released under an open source license in 2018 [12]. It is used to pre-train deep two-dimensional representations from unlabeled text with general conditioning of both left and right contexts in all classes.

The representation of the BERT system allows it to be used as a basis for measuring the similarity of sentences in natural languages. Here, we use a measure of the distance between the text attachments of the sentences being compared. Typically, the text measure or text generation quality measure is a function:

$$f(x, \hat{x}) \in \mathbb{R} \quad (4)$$

Where:

$x \rightarrow < x_1, x_2, \dots, x_k >$ is the vectorized representation of the sample proposal

$\hat{x} \rightarrow < \hat{x}_1, \hat{x}_2, \dots, \hat{x}_l >$ is the vectorized representation of the candidate proposal

A good measure must reflect the person's judgment as accurately as possible. That is, it must show a high correlation with the person's assessment results.

Compared to the LSTM, GRU, and BiLSTM models, the BERT model has a great advantage because it is pre-trained on a large amount of Vietnamese text data, allowing it to learn rich and deep linguistic features. When applying BERT, the model can take advantage of the knowledge learned about the Vietnamese language into the problem of fake news detection.

However, BERT also has the disadvantage of having a large model size and requiring a lot of computing resources. Therefore, training and using BERT in real applications requires more powerful hardware than models based on LSTM, GRU, and BiLSTM.

4 Experimental Process

The system has been experimented and evaluated through Vietnamese data. The experiment process is performed as follows:

4.1 Data Collection

Data is collected automatically using the Crawler tool. The tool has supported the collection of more than 2,000 articles on two topics: politics and medical. For each topic, data is divided into two types of news: real news and fake news. Real news sources are mainly collected from official, reputable newspaper sites in Vietnam such as: *nhandan.vn*, *dantri.com.vn*, *thanhnien.vn*, ... Unofficial news sources system, or contains a lot of fake and distorted news collected at sites such as: *quanlambao.blogspot.com*, *viettan.org*, *binhluan.biz*, ... All downloaded news go through manual selection and labeling to ensure reliability. The news will be labeled 0 if it is real news and will be labeled 1 if it is fake or distorted news.

Statistics of the collected data and labeling according to the number of news and number of sentences are as follows:

Table 1
Statistics of collected data and labeling

Topic	Category	Training		Testing		Total	
		News	Sentence	News	Sentence	News	Sentence
Politics	Real news	419	14,992	99	2,796	517	17,788
	Fake news	396	25,581	110	4,407	506	29,988
Medical	Real news	473	24,086	105	3,214	579	27,300
	Fake news	397	17,712	107	4,552	504	22,264
Total		**1,658**	**82,371**	**421**	**14,969**	**2,106**	**97,340**

The process of selecting data and labeling ensures certain properties such as balance between the amount of fake news and real news to avoid the phenomenon of imbalanced data. This data is divided into a training dataset accounting for about 80% and a testing dataset used only to evaluate the model independently and not participating in the training process accounting for about 20%.

For each collected news, they have a structure including the following attributes:

Table 2

Attribute structure of a news

№	**Field name**	**Types**	**Description**
1	Category	Text	Name of news category (example: news about the medical field, about the political field,...)
2	Source	Text	Source (link) where news is shared
3	Release time	Date	Posting time
4	Topic	Text	The title of the news
5	Content	Text	Content of news
6	Label	Integer	0: real news, 1: fake news

Each news after processing will be stored as follows:

```
loai:        "y_te"
nguon:       "viettan.org"
link:        "https://viettan.org/ty-le-binh-phuc-tu-covid-o-viet-nam-qua-thap/"
tg_dang_tin: "2021-09-07 00:00:00"
tieu_de:     "Tỷ lệ bình phục từ Covid ở Việt Nam quá thấp?"
noi_dung:    "Một điều ngạc nhiên là trong trận dịch này, tỷ lệ người nhiễm và bình phục ở Việt Nam chỉ 55%,
              tính đến nay (4/9/2021), đã lên đến 12.793 người. Ở Thái Lan, con số tử vong là 12.631 người,
              covid ở Việt Nam xem ra có nguy cơ chết cao hơn Thái Lan rất nhiều. Tuy nhiên, còn một con số
              chính xác, có thể phản ảnh phác đồ điều trị hay/và năng lực và chất lượng của hệ thống y tế của
              và số người bình phục là chỉ 55% (n = 282.516). Con số 55% này rất thấp nếu so với tất cả các n
              chúng ta tính toán tỷ lệ tử vong trên số ca nhiễm có lẽ không tốt mấy. Cách khác là tính trên
              cách tính tỷ lệ tử vong thực tế hơn là: số ca tử vong / (số ca tử vong + số ca bình phục) Biểu
              cả các nước trong vùng. Số ca nhiễm ở Thái Lan (1,26 triệu) và Mã Lai (1,82 triệu), nhưng hai
              Chỉ 1,1%. Phân tích như thế mới thấy Việt Nam bị nặng nề nhứt. Tỷ lệ bình phục là một chỉ số phả
              số này có thể nói lên rằng hệ thống y tế của Việt Nam có vấn đề trong việc đáp ứng đại dịch. Kh
              vong covid là có các bệnh đi kèm. Bệnh đi kèm phổ biến nhứt là: • viêm phổi • bệnh đường hô hấ
              nhớ (dementia) Điều này có nghĩa thực tế là rất khó xác định Covid-19 là nguyên nhân trực tiếp
              đoạn cuối và sau đó bị Covid-19 và chết, thì theo quy định hiện hành bác sĩ phải ghi là "Chết
              là thủ phạm bồi thêm. Tôi hay ví von là bệnh nền như là súng đã lên đạn, và Covid-19 là nó bóp
              covid," chớ tôi không viết "tử vong vì covid". Thật thú vị, hôm kia một quan chức y tế Úc mới
              virus" ("die with virus"), chớ không phải "chết vì virus" ("die from virus"). Tôi nghĩ đó là m
              ngoái. Từ nay trở đi, chúng ta nên học cách "live and die with the virus" – "sống và chết cùng
              Đảng Việt Tân"
nhan:        1
```

Figure 5

Example of news in the data set

With the attribute structure of such an organized news, it not only helps to detect fake news based on content but also serves for research based on other analytical techniques (traceability, time of posting,...).

4.2 Data Preprocessing

The data preprocessing process is based on tools that perform the following steps: data cleaning (analysis to only retain necessary content such as titles, text content,…); word separation (using the function *ViTokenizer.tokenize()* of the library *pyvi* to support word separation and word classification for Vietnamese using the algorithm CRFs - Conditional Randomields); remove stopwords (to reduce the size of the data to speed up model training without affecting the style and meaning of the sentence); data standardization (select 10,000 vocabulary words out of a total of more than ~14,000 words of the collected data set, each word will be represented by an integer and arranged accordingly in descending order of number of times occurrence of words in the data set).

4.3 Word Embedding

To implement Word Embedding for fake news classification application, this study used a technique of *Word2Vec*, the CBOW (Continuous Bag of Words) model.

4.3.1 The Preparation of Contextual Data

Based on the sentences of the standardized news set, build a set of contexts and a set of target words to train the CBOW model with 4-grams (four context words around a target word).

Input: set of standardized news (in sentences)

Output: the context set has size *(T, 4)* and the corresponding target set has size *T*

The program segment that prepares training data for the CBOW model includes a set of context words corresponding to the main word set:

```
context_length = self.window_size * 2 # Initialization self.window_size = 2
X = []
y = []
for words in tqdm(sequences_list): # Browse all sentence sets
    sentence_length = len(words)
    for idx, w in enumerate(words): # Browse all the words of the sentence
        context_words = []  # Initialize the surrounding sliding frame from w
        start = idx - self.window_size
        end = idx + self.window_size + 1
        context = []
        for i in range(start, end):
            if (0 <= i < sentence_length) and (i != idx):
                context.append(words[i])
```

```
        context_words.append(context) #Surrounding words as context
        sample_x = pad_sequences(context_words, maxlen=context_length,
                                 padding='post', truncating='post')
        X.append(sample_x[0])
        y.append(w)    # add w as target word
X = np.array(X)  # context set (X)
y = np.array(y)  # target word set (Y)
np.save(self.x_cbow_path, X)
np.save(self.y_cbow_path, y)
```

4.3.2 Building a CBOW Neural Network Model

The CBOW neural network model is installed according to the following parameters:

- Dictionary size (V): 10.000
- Vector space size (N) is also the number of neurons in the hidden layer: 128
- Number of context words (C) surrounding the target word: 4
- Learning rate: 0.01
- Number of training loops (epochs): 500

In addition, the model training process uses the early automatic stopping technique (Early Stopping), the number of consecutive training epochs as the non-decreasing loss function value is set to 6. The CBOW rneural network structure is as follows:

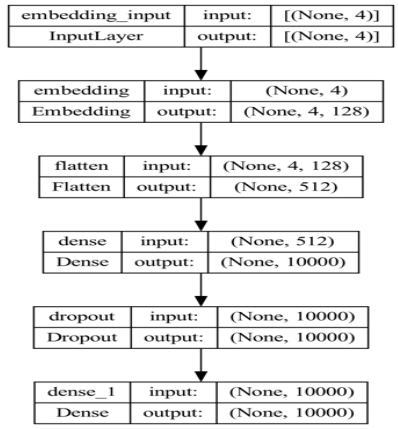

Figure 6
Network model CBOW

4.3.3 Results after Word Embedding

After training the CBOW network model with the data set of contexts and target words of fake news dataset, a matrix W of size V x N will be extracted. Where N is the number of neurons of the hidden layer in CBOW and also is the size of the vector to represent each word.

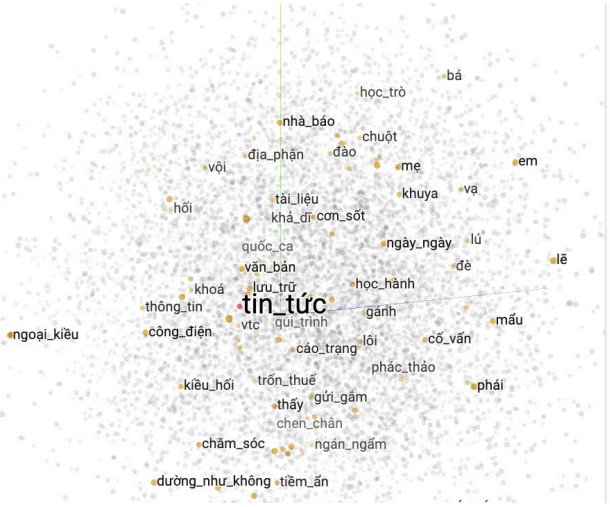

Figure 7
Word Embedding representation of fake news classification data

Figure 7 provides a visual representation of co-occurrences and neighboring words based on calculations done with CBOW. With this representation, we can see which words often appear with a given word.

4.4 BERT Model

4.4.1 Configure Training Parameters

Configuring the training parameters for the models is done as follows:

- Each news is represented as a set of vectors. The maximum number of words in each news is limited to 512 words. This number was chosen based on actual observations of the average length of news after data cleaning. Limiting this number of words helps ensure a fixed size before feeding data into training models.
- After performing Word Embedding, each word in the news is represented as a 128-dimensional vector, so the size of a news representation will be 512 x 128, in which 512 represents the maximum number of words in each news and 128 is the size of each word representation vector.

The program segment performs training for the models:

```
def train(self):
    # Read training data
    X_train, y_train= self.get_data(news_path=self.train_news_path)
    # Read the model
    model = self.get_model()
    # stop when 6 consecutive times there is no loss reduction
    early_stop = EarlyStopping(monitor='val_loss', patience=6, verbose=1)
    # Set up model auto-save
    model_checkpoint = ModelCheckpoint(f'{self.weights_folder}/{
model.model_name}_best.h5', monitor='val_loss', verbose=1)
    # Model training
    model.fit(X_train, y_train, validation_split=0.2, epochs=1000, callbacks=[early_stop,
model_checkpoint])
    # Save the model after training
    model.save(f'{self.weights_folder}/ {self.rnn_cls.model_name}.h5')
```

The training process of the models is specifically measured through the following charts:

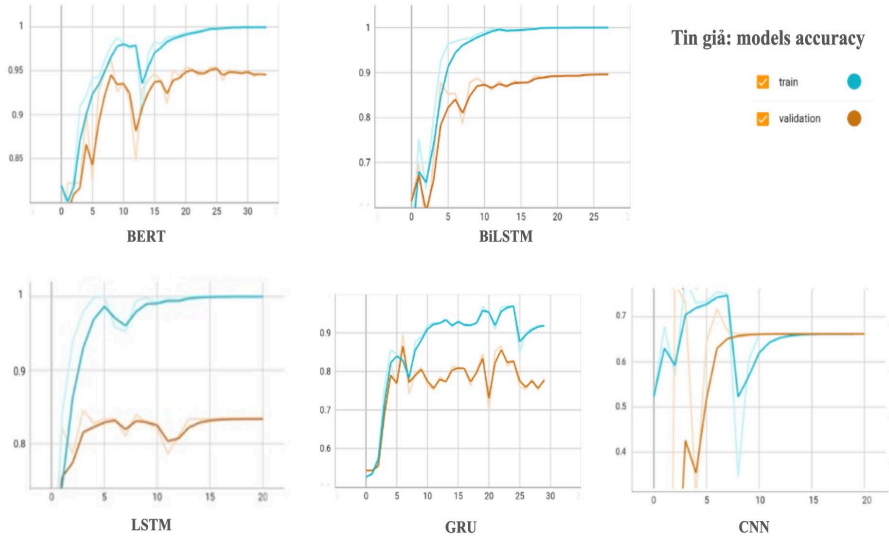

Figure 8
Model evaluation chart during training process

Figure 8 is the result illustrating the historical data returned, with the vertical axis being ***Cross Entropy*** (val_loss value) and the horizontal axis being ***Epoch*** (number of iterations). The cross-entropy loss for the training dataset is accessed via the 'loss' key. The loss on the validation dataset is accessed via the 'val_loss' key on the history object's history property.

5 Analysis and Evaluation of Experimental Results

To evaluate the results of detecting fake or real news, the system was tested on a testing dataset of 421 news, including 204 real news and 217 fake news.

Testing is performed through the model evaluation program with the following test data set:

Input: Testing dataset (X_test, y_test)

Output: Evaluation results

```
def evaluate(self):
    # Read test data set
    X_test, y_test =
                self.get_data(news_path=self.test_news_path)
    # Read the model
    model = self.get_model()
    # Apply the trained weights to the model
    model.load_weights(f'{self.weights_folder}/
                              {self.rnn_cls.model_name}.h5')
    # Evaluate the model
    ret = model.evaluate(test_ds,verbose=1,return_dict=True)
return ret
```

The accuracy of spy detection is calculated through measurements such as Accuracy, Recall, Precision, F1-score, ROC-AUC and the specific results are as follows:

Table 3
Results achieved through models

Model	Accuracy	Recall	Precision	F1-score	ROC-AUC
BERT	**0.9455**	**0.9507**	**0.9369**	**0.9438**	**0.9466**
BiLSTM	0.9097	0.8981	0.9158	0.9069	0.9022
LSTM	0.8147	0.7962	0.8276	0.8116	0.8100
GRU	0.7720	0.7476	0.7778	0.7624	0.7577
CNN	0.7083	0.6716	0.7098	0.6903	0.6864

Table 3 shows that applying the proposed solution based on CBOW and BERT gives better results than other methods such as CNN, GRU, LSTM and BiLSTM.

Conclusions

This paper has proposed a solution to support fake news detection, based on the combination of CBOW and BERT. This solution allows calculating the contextual characteristics of the language, when vectorizing text, to improve accuracy, when analyzing text. In addition, the use of the BERT model has contributed to improving the quality and speed of calculations on the training data set standardized by

CBOW. Experimental results of the new solution, compared to other methods, such as CNN, GRN, LSTM and BiLSTM on the same dataset, shows that it gives better results.

This architecture and model is completely applicable to other languages. However, it may be necessary to experiment, to adjust the parameters of the CBOW and BERT network, to find the optimal parameters. Adjusting the parameters can depend on the context the word window, the size of the dictionary, the dimensionality of the word vector and the parameters of the BERT network.

In the future, we will continue our research and try to apply it to other fields. That is, other than politics and pandemics. We will also experiment with new methods to update the training data set and detect fake news within a multilingual context.

Acknowledgement

This research was funded by the Ministry of Education and Training (Vietnam) through the project code B2022-DNA-17.

References

[1] J. Strömbäck et al., News media trust and its impact on media use: toward a framework for future research, Annals of the International Communication Association, Volume 44, Issue 2, 2020 https://doi.org/10.1080/23808985.2020.1755338

[2] M. Amjad et al., Survey of Fake News Datasets and Detection Methods in European and Asian Languages, Acta Polytechnica Hungarica, Vol. 19, No. 10, pp. 185-204, 2022; https://doi.org/10.12700/APH.19.10.2022.10.11

[2] T. H. Vo et al., Automatic detection of fake news: Achievements and challenges, Journal of Science and Technology, Vol. 20, No. 3, pp. 71-78, 2022; https://doi.org/10.15587/1729-4061.2022.265317

[4] J. Alghamdi et al., A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection, Journal of Information, Volume 13, No. 12, 2022; https://doi.org/10.3390/info13120576

[5] V. M. Kresnakova et al., Deep learning methods for Fake News detection, Proceedings of 19th International Symposium on Computational Intelligence and Informatics, published by IEEE, pp. 143-148, 2020 https://doi.org/10.1109/CINTI-MACRo49179.2019.9105317

[6] I. K. Sastrawan et al., Detection of fake news using deep learning CNN–RNN based methods, ICT Express,Volume 8, Issue 3, pp. 396-408, 2022 https://doi.org/10.1016/j.icte.2021.10.003

[7] C. S. Rao et al., Effective Fake News Classification Based on Lightweight RNN with NLP, Annals of Data Science, 2024 https://doi.org/10.1007/s40745-023-00506-z

[8] S. R. Tanuku, Novel Approach to Capture Fake News Classification Using LSTM and GRU Networks, Proceedings of 2022 International Conference on Futuristic Technologies (INCOFT), IEEE, 2022 https://doi.org/10.1109/INCOFT55651.2022.10094467

[9] S. H. Nithya and A. Sahayadhas, Automated Fake News Detection by LSTM Enabled with Optimal Feature Selection, Journal of Information & Knowledge Management, Volume 21, No. 03, 2022 https://doi.org/10.1142/S0219649222500368

[10] D. S. Asudani et al., Impact of word embedding models on text analytics in deep learning environment: a review, Artificial Intelligence Review, Volume 56, pp. 10345–10425, 2023; https://doi.org/10.1007/s10462-023-10419-1

[11] B. Li et al., Scaling Word2Vec on Big Corpus, Data Science Engineering, Volume 4, pp. 157-175, 2019; https://doi.org/10.1007/s41019-019-0096-6

[12] A. Özçift, Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish, Journal for Control, Measurement, Electronics, Computing and Communications, Volume 62, Issue 2, pp. 226-238, 2021 https://doi.org/10.1080/00051144.2021.1922150