A study on deep learning for Vietnamese text classification

Nguyen Thi Hien[1*], Bui Thi Thoa[1] , Luong Nguyen Hoang Hoa[2]

[1]Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Hanoi, Vietnam;
[2]Ministry of Public Security, 44 Yet Kieu, Hoan Kiem, Hanoi, Vietnam.
[*]Corresponding author: hiennt@lqdtu.edu.vn
Received 12 Jan. 2024; Revised 4 Mar. 2024; Accepted 10 May 2024; Published 20 May 2024.
DOI: https://doi.org/10.54939/1859-1043.j.mst.95.2024.85-94

ABSTRACT

Text categorization aims to automatically assign given text passages or documents to predetermined categories or subjects. Despite the wide array of techniques employed in classifying English text, there remains a dearth of research on Vietnamese text classification. This paper introduces a novel approach utilizing a Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) with a deep network structure for Vietnamese text classification. Our findings demonstrate a substantial improvement in classification accuracy when applying deep learning techniques to two Vietnamese news corpus datasets. This study contributes to the advancement of Vietnamese text classification by introducing and demonstrating the efficacy of LSTM and CNN with a deeper network structure. The results offer valuable insights for researchers and practitioners working on text categorization in the Vietnamese language.

Keywords: Deep learning; Text classification; LSTM; CNN.

1. INTRODUCTION

Text classification is a natural language processing problem that assigns labels to textual units like sentences, queries, paragraphs, and documents. It has various applications like question answering, spam detection, sentiment analysis, news categorization, and user intent classification. Extracting insights from text is challenging due to its unstructured nature.

Advances in machine learning have made automatic text categorization systems more effective. Machine learning models, often based on two-step procedures, extract hand-crafted features from documents and use them to predict predictions. Popular algorithms include Naïve Bayes (NB), Support Vector Machine (SVM), hidden Markov model (HMM), gradient boosting trees, and Random Forests (RF). However, this approach has limitations, including tedious feature engineering, difficulty generalizing to new tasks, and pre-defined feature templates, which limit the use of large training data.

Recently, there has been a growing interest in exploring deep learning models to address some of the fundamental limitations of machine learning. This paper reviews deep learning models developed for text classification tasks. The models are categorized based on their neural network architectures, including Recurrent Neural Networks, Convolutional Neural Networks, Attention and Transformers.

Besides, the Vietnamese language faces a growing need for an effective automated classification system due to its importance. Previous research has focused on classical machine learning classifiers and small datasets. The researchers use nine robust deep neural network-based classifiers tested on large datasets, achieving high accuracy for single and multi-label categorization tasks.

Furthermore, the stated accuracy of such solutions has a lot of potential for

improvement. We construct two robust deep learning based classifiers that have been tested on large datasets and produce excellent accuracy for Vietnamese text classification tasks.

The rest of the paper is organized as follows: section 2 discusses related work, section 3 then presents methodologies for Vietnamese text classification, section 4 provides a detailed overview of the dataset utilized, gives the results of experiments and section 5 reports our conclusions and future works.

2. RELATED WORKS

Text classification involves the assignment of documents to predefined categories or classes. This is not a new issue; since the 1800s, humans have used knowledge engineering approaches to automatically categorize texts. Previously, similar approaches were utilized for manual categorization.

These days, with the popularity of deep learning and machine learning, both may be utilized for text categorization, among other classification tasks. Text classification is viewed as an issue of classifying documents in both machine learning and deep learning. A predetermined corpus is used by automatic text categorization algorithms for training and learning. We choose features from the corpus for every category. Next, based on these traits, a mathematical model, also known as a classifier, is used to evaluate the degree of similarity between texts and categorize them.

There are different methods to handle this challenge. Most commonly used are Naive Bayes (NB), Logistic Regression (LR) [1], Decision Tree (DT) [2], and Support Vector Machine (SVM) [3], and other study which are considered top-notch for English processing. These methods show how machine learning techniques are useful for solving text classification problems.

Recent approaches based on deep learning have showcased the capability for scalable classification on extensive engineering document datasets [4]. These approaches concentrated on textual information and employed various Natural Language Processing (NLP) techniques to create automated classifiers, including Convolutional Neural Network (CNN) [5-7], Recurrent Neural Networks (RNN) [8], Long Short-term Memory (LSTM) networks [9], and specific pre-trained models.

For Vietnamese text classification, one of the challenges faced is that a space character does not always represent word boundaries. Vietnamese articles comprise various types of words, including single words, compound words, and duplicative words. Additionally, there are instances of fortuitous concurrence words. Vietnamese words are typically formed by special linguistic units known as morpho-syllables. Each unit can be a morpheme, a word, or neither, and the process of identifying these units is referred to as word segmentation. An example of word segmentation in a Vietnamese sentence is provided below:

- Sentence: bảo lãnh bằng chứng khoán.
- Case 1: bảo_lãnh bằng_chứng khoán.
- Case 2: bảo_lãnh bằng chứng_khoán.

This is what makes the task of classifying Vietnamese text more challenging. For instance, in case 1, the text might be classified as legal, while in case 2, it could be categorized as financial text.

Several studies on Vietnamese text classification have achieved good accuracy by employing methods such as machine learning techniques [10, 11], Statistical N-Gram Language Modeling [12] and deep learning [9]. These studies use different data sets and most likely have less than 90% accuracy. In this research, we utilize the same corpus as [12] to evaluate the effectiveness of both machine learning and deep learning algorithms for the task of Vietnamese text classification.

3. PROPOSED METHODS

Our research consisted of studying different approaches to Vietnamese document classification, evaluating those approaches, and finally comparing them in order to know the best performing method. The text classification process simulation is shown in Figure 1, 2 below:

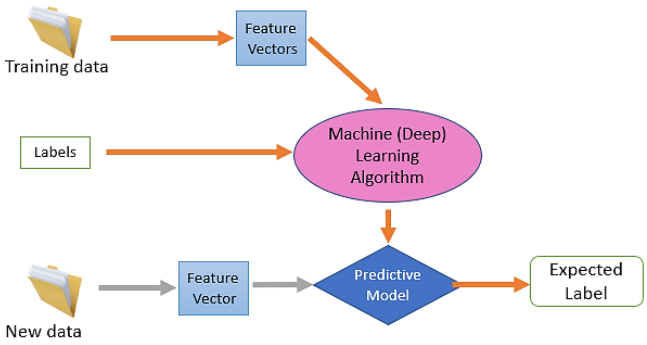

Figure 1. *The text classification process.*

There are three primary parts to the text classification process. Preprocessing, which involves cleaning the input text, is the first step. Text feature extraction makes up the second part, while text classification model makes up the last. Each part's specifics are as follows.

3.1. Pre-processing

Many stopwords, misspellings, slang, etc, are unnecessary words contained in most text and document data sets. Noise and extra characteristics can negatively impact system performance for several algorithms, particularly statistical and probabilistic learning algorithms. We discuss shortly a few text cleaning and pre-processing strategies and methodologies data sets in this section.

3.1.1. Word tokenization

To efficiently tokenize words, a robust document classification method requires a strong word segmentation approach in Vietnamese. So, vnTokenizer [13] is utilized to tokenize the text documents into words or tokens.

3.1.2. Stop-words removal

Before extracting document features, an important task is to eliminate stop-words and perform stemming.

Stop-words are words that are widely used but contribute very little useful information. Stop-words will be excluded during the text processing. For example, in Vietnamese, the stop words include: "là", "và", "của", etc. To facilitate this, a stop-word list for Vietnamese is used. It includes nearly 2000 Vietnamese stop-words.

3.2. Features extraction

For this task, we use Bag of Words and Term Frequency-Inverse Document Frequency to transform documents within our corpus into vectors.

3.2.1. Bag of Words

Bag of Words (BOW) is a method for extracting features from textual data. This method requires a predefined set of words, known as the bag of words. The features of a text are represented through the presence of the words from that text in the bag of words. For example, let's consider a bag of words as:

BOW vocabulary:{“happy”:0, ”today”:1, ”beautiful”:2, ”not”:3, “day”:4}.

Now, two sentences, "Today I am not happy, but he is happy" and "Today is a beautiful day" would be represented as [2, 1, 0, 1, 0] and [0, 1, 1, 0, 1], respectively.

BOW is a simple and effective method of text representation widely used in machine learning tasks such as text classification. However, BOW has the drawback of losing information about word order, and a large dictionary size poses challenges in terms of memory usage and computational complexity.

3.2.2. Term Frequency-Inverse Document Frequency (TF-IDF)

- Term Frequency (TF)

The term is the frequency measure of a word w in a document (text) d. It is equal to the number of instances of word w in document d divided by the total number of words in document d. Term frequency serves as a metric to determine a word’s occurrence in a document as compared to the total number of words in a document. The denominator is always the same.

$$Term\ Frequency = \frac{number\ of\ instances\ of\ word\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d} \quad (1)$$

- Inverse Document Frequency (IDF)

This parameter gives a numeric value of the importance of a word. Inverse Document frequency of word w is defined as the total number of documents (N) in a text corpus D, divided by the number of documents containing w.

$$IDF = log\frac{number\ of\ documents\ (N)\ in\ a\ text\ corpus\ D}{number\ of\ documents\ containing\ w} \quad (2)$$

The product of TF and IDF is the TF-IDF. TF-IDF is usually one of the best metrics to determine if a term is significant to a text. It represents the importance of a word in a particular document.

After applying TF-IDF to represent the documents in a numerical format, then Singular Value Decomposition (SVD) is employed for dimensionality reduction and feature extraction. This combination of techniques helps in capturing the essential patterns and relationships within the TF-IDF-weighted matrix, enhancing the efficiency of subsequent analysis or machine learning models.

SVD is a unique matrix decomposition that exists for every complex-valued matrix $X \in \mathbb{C}^{n\times m}$ [14]:

$$X = U \sum V^T$$

where $U \in C^{n \times n}$ and $V \in C^{m \times m}$ are unitary matrices with orthonormal columns, and $\sum R^{n \times m}$ is a matrix with real, nonnegative entries on the diagonal and zeros off the diagonal.

3.3. Classifier models

We apply two deep learning methods are Long Short Term Memory (LSTM), and Convolutional Neural Network (CNN). The algorithms will be presented below:

3.3.1. Long Short Term Memory (LSTM)

LSTM Networks, designed by Hochreiter and Schmidhuber [15], are deep learning, sequential neural networks that can effectively handle multiclass classification problems, particularly with sequential data like text or time series. They store separate memory cells, implement three layers: (1) inputs gate, (2) forget gate, and (3) output gate [14].

Each LSTM unit, can be seen in Figure 3 has a memory cell, and the states at time t are represented as c_t. Reading and modifying are controlled by the sigmoid gate and affect the input gate i_t, forget gate f_t and output gate o_t.

A step-by-step calculation of the LSTM cell and its gates are provided below:

- The first step, the forget gate:

$$f_t = \sigma(x_t * U_f + h_{t-1} * W_t)$$

- The input gate is provided below:

$$i_t = \sigma(x_t * U_i + h_{t-1} * W_i)$$
$$N_t = \tanh(x_t * U_c + h_{t-1} * W_c$$

- Cell state:

$$C_t = f_t * C_{t-1} + i_t * N_t$$

- Output gate is provided by the formula below:

$$o_t = \sigma(x_t * U_0 + h_{t-1} * W_0)$$
$$h_t = o_t * \tanh(C_t)$$

Where x_t refers to input to the current timestamp t, U_f is weight associated with the input, h_{t-1} is the hidden state of the previous timestamp and W_f is the weight matrix associated with the hidden state, σ indicates the function of the logistics sigmoid.

In Vietnamese Text Classification task, we create a LSTM network as follows figure 3:

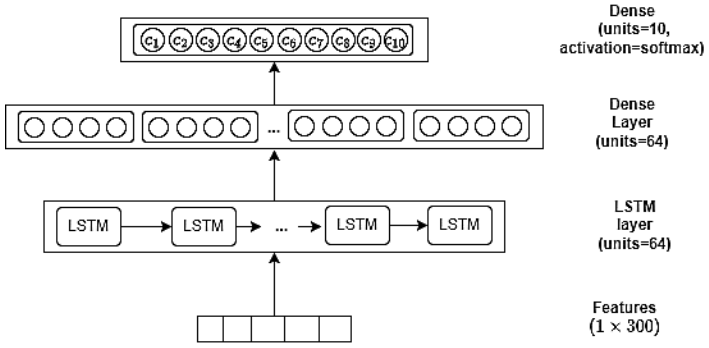

Figure 2. *LSTM network architecture for Vietnamese text classification.*

3.3.2. Convolutional Neural Network

A common type of neural network used for tasks involving natural language processing is the Convolutional Neural Network (CNN). It works particularly well when dealing with data sequences, like text. CNN can be used to perform language modeling, machine translation, and text classification.

A non-linear activation function is used to pass the filter output from each convolutional layer, which applies a series of filters to the input data before moving on to the next layer. By making the network more complex through the use of the activation function, the CNN is able to learn more intricate patterns and features. One or more "fully-connected" layers are frequently employed following the output of the convolutional layers to provide predictions or assessments based on the patterns and features that the CNN identified. The criteria for text classification closely resemble those of image classification, with the key distinction being the utilization of a matrix of word vectors instead of pixel values. The authors developed a Convolutional Neural Network model with two convolutional layers and an activation function of RELU for the Vietnamese Text Classification challenge. Adam was utilized to tune the network. The input layer is the feature vector after the feature selection phase and the output layer is the label vector of the documents. Details about the CNN model used for Vietnamese text classification are described in figure 4 below.

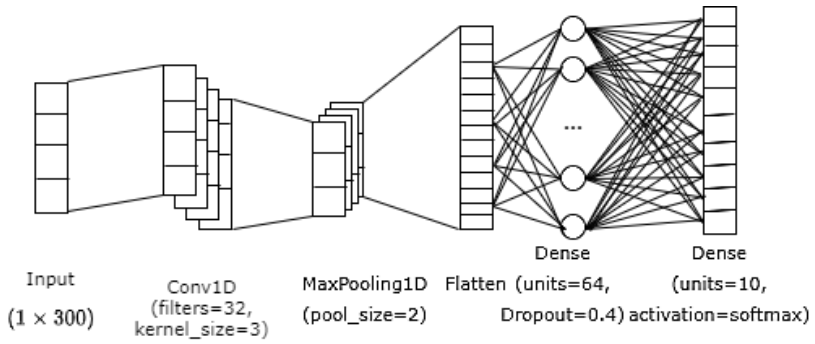

Figure 3. *CNN model architecture used for Vietnamese text classification.*

4. EXPERIMENTS AND RESULTS

In this study, we conducted evaluation experiments on the dataset named "A Large-scale Vietnamese News Text Classification Corpus" [12]. This dataset based on the four largest circulation Vietnamese online newspapers: VnExpress, TuoiTre Online, Thanh Nien Online, Nguoi Lao Dong Online. The online documents obtained through web crawling are automatically altered (e.g., removing HTML tags, spelling normalization) by Teleport5. Subsequently, linguists manually review and correct the text documents that have been misclassified into predefined topics. This dataset comprises two levels.

- Level 1: Level 1 includes some top categories from the above popular news websites. This contains about 33,759 documents for training and 50,373 documents for testing.

- Level 2: Level 2 includes the topics that are child topics in level 1. Level 2 contains about 14375 documents for training and 12076 documents for testing. The categorization at Level 1 is quite ambiguous, and there is a need to identify specific topics for text classification. Both levels of the dataset are utilized to conduct experiments with our classification models. Corpus level 1 and 2 are described by table 1 and 2 and as follows.

In all experiments, we employ accuracy to evaluate the classification models. The accuracy is defined as below:

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \quad (3)$$

The experiments were conducted using two text representation methods, Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). In this experiment, we compare the results of CNN and LSTM with 3 typical machine learning algorithms (LR, DT and SVM) for the text classification problem. With machine learning algorithms such as LR and DT, default parameters were used during the experimentation. For SVM, we employed the SVC type with C=1. On the other hand, for deep learning algorithms, due to limited data and a small input representation vector, so we chose to use more straightforward network architectures. CNNs network consists of one convolutional layer followed by MaxPooling, Flatten, and Dense layers. For the LSTM network was designed a model with one LSTM layer followed by Dense and Dropout layers. The experimental results are in table 3 and 4 below:

The results of experimenting with machine learning algorithms, including Logistic Regression, Decision Trees, and Support Vector Machines, as well as deep learning algorithms such as Convolutional Neural Networks and Long Short Term Memory, were presented. The results in Tables 3 and 4 show that with the TF-IDF features extraction technique, machine learning methods have better results, but with deep learning methods, the results decrease a bit, although not significantly. Table 3 results also show that the use of deep learning techniques is better than machine learning techniques. This also shows more clearly that with the first data set, the division into unclear categories (it is possible to confuse one category with another), deep learning techniques give much better results than machine learning techniques.

Table 1. *Detail of level 1 dataset.*

No	Topic	Train	Test
1	polities-society	5,129	7,567
2	life	2,159	2,036
3	Science& technology	1,820	2,096
4	business	2,552	5,276
5	heath	3,384	5,417
6	law	3,868	3,788
7	world news	2,898	6,716
8	sports	5,298	6,667
9	culture	3,080	6,250
10	informatics	2,481	4,560
	Summary	**5,129**	**7,567**

Table 2. *Detail of level 2 dataset.*

No	Topic	Train	Test
1	music	900	813
2	eating and drinking	265	400
3	real property & technology	246	282
4	football	1,857	1,464
5	stock	382	320
6	bird flu - influenze	510	381
7	the life in the world	729	405
8	studying abroad	682	394
9	tourist	582	565
10	WTO	208	191
11	family	213	280
12	computer entertainment	825	707
13	education	821	707
14	sex	343	268
15	hackers and viruses	355	319
16	criminal	155	196
17	life space	134	58
18	international business	571	559
19	Beauty	776	735
20	lifestyle	223	214
21	shopping	187	84
22	fine arts	193	144
23	stage and screen	1,117	1,030
24	new computer products	770	595
25	tennis	588	283
26	young world	331	380
27	fashion	412	302
	Summary	**14,375**	**12,076**

Table 3. *The accuracy of algorithms with the BOW representation method.*

	LR	DT	SVM	LSTM	CNN
Data Level 1	92.29	78.31	91.85	**94.52**	93.75
Data Level 2	87.79	68.13	86.08	**89.35**	88.22

Table 4. *The accuracy of algorithms with the TF-IDF representation method.*

	LR	DT	SVM	LSTM	CNN
Data Level 1	92.77	83.60	93.60	93.75	93.59
Data Level 2	88.20	73.82	89.35	89.43	88.33

5. CONCLUSIONS AND FUTURE WORKS

In this study, we conducted research and built experiments on utilizing deep learning models for the task of Vietnamese text classification. Throughout our research, we performed experiments and compared results among different approaches to gain insight

and accurate understanding of the performance of deep learning and machine learning classification algorithms, specifically for the task of Vietnamese text classification.

Additionally, several errors occur in these approaches for Vietnamese text classification. They can be described as follows:

- Text representation methods are relatively simple (BOW and TF-IDF).
- The corpus contains ambiguities between two or many topics.
- The segmentation is limited by a third-party library.

In the future, we could enhance the accuracy of deep learning algorithms for this problem, address the disadvantages of preprocessing phases, and integrate more semantic and contextual features in this text classification problem for Vietnamese.

Acknowledgement: *This research is funded by the project CNC.2021.B05.03.*

REFERENCES

[1]. P. Komarek, *"Logistic regression for data mining and high-dimensional classification",* Carnegie Mellon University, (2004).

[2]. M. N. M. S. a. A. H. O. W. N. H. W. Mohamed, *"A comparative study of Reduced Error Pruning method in decision tree algorithms",* in IEEE International Conference on Control System, Computing and Engineering, Penang, (2012).

[3]. C. &. V. V. Cortes, *"Support-vector networks,"* Machine learning, vol. 20, pp. 273-297, (1995).

[4]. L. A. a. F. Tietze, In: World Patent Information , (2018).

[5]. Y. Kim, *"Convolutional neural networks for sentence classification,"* arXiv preprint arXiv:1408.5882, (2014).

[6]. X. J. Z. a. Y. L. Zhang, *"Character-level convolutional networks for text classification,"* Advances in neural information processing systems, vol. 28, (2015).

[7]. S. a. C. S. Moriya, *"Transfer learning method for very deep CNN for text classification and methods for its evaluation,"* 2018 IEEE 42nd annual computer software and applications (COMPSAC), (2018).

[8]. S. L. X. K. L. a. J. Z. Lai, *"Recurrent convolutional neural networks for text classification,"* in The AAAI conference on artificial intelligence, (2015).

[9]. W. K. D. P. R. a. R. F. M. Sari, *"Text classification using long short-term memory,"* in International Conference on Electrical Engineering and Computer Science (ICECOS), (2019).

[10].T. H. N. H. N. D. L. T. a. V. T. N. Nguyen, *"A hybrid feature selection method for Vietnamese text classification"* in IEEE Seventh International Conference on Knowledge and Systems Engineering (KSE), (2015).

[11].H. T. N. D.-T. D. Q. T. a. H. X. H. Huynh, *"Vietnamese text classification with textrank and jaccard similarity coefficient,"* Adv. Sci. Technol. Eng. Syst 5, vol. 5, no. 6, (2020).

[12].V. C. D. D. D. L. N. N. &. N. H. Q. Hoang, *"A comparative study on vietnamese text classification methods,"* in IEEE international conference on research, innovation and vision for the future, (2007).

[13].N. M. D. B. N. N. V. D. &. N. T. D. Le, *"VNLP: an open source framework for Vietnamese natural language processing,"* in Proceedings of the 4th Symposium on Information and Communication Technology, (2013).

[14].N. B. S. L. &. N. K. J. Benjamin Erichson, *"Compressed singular value decomposition for image and video processing,"* in Proceedings of the IEEE International Conference on Computer Vision Workshops, (2017).

[15].S. &. S. J. Hochreiter, *"Long short-term memory. Neural computation,"* vol. 9, no. 8, pp. 1735-1780, (1997).

[16].A. Graves, *"Generating sequences with recurrent neural networks.,"* arXiv preprint arXiv:1308.0850, (2013).
[17].W. &. J. W. Dai, *"A mapreduce implementation of C4. 5 decision tree algorithm. International journal of database theory and application,"* vol. 7, no. 1, pp. 49-60, (2014).
[18]. H. N. a. N. T. M. A. Phat, *"Vietnamese text classification algorithm using long short term memory and Word2Vec,"* Информатика и автоматизация, vol. 19, no. 6, pp. 1255-1279, (2020).

TÓM TẮT

Nghiên cứu kỹ thuật học sâu cho bài toán phân lớp dữ liệu tiếng Việt

Phân loại văn bản nhằm mục đích tự động gán các đoạn văn bản hoặc tài liệu nhất định thuộc vào các danh mục hoặc chủ đề được xác định trước. Mặc dù có rất nhiều kỹ thuật được sử dụng để phân loại văn bản tiếng Anh nhưng vẫn còn thiếu các nghiên cứu về phân loại văn bản tiếng Việt. Bài viết này giới thiệu một cách tiếp cận mới sử dụng Bộ nhớ ngắn hạn dài (LSTM) và Mạng tích chập (CNN) với cấu trúc mạng nơ-ron sâu để phân loại văn bản tiếng Việt. Phát hiện của chúng tôi chứng minh sự cải thiện đáng kể về độ chính xác trong phân loại khi áp dụng các kỹ thuật học sâu cho hai tập dữ liệu tin tức tiếng Việt. Nghiên cứu này góp phần thúc đẩy sự cải tiến của phân loại văn bản tiếng Việt bằng cách giới thiệu và chứng minh tính hiệu quả của LSTM và CNN với cấu trúc mạng sâu. Kết quả mang lại những hiểu biết sâu sắc có giá trị cho các nhà nghiên cứu và thực hành nghiên cứu về phân loại văn bản trong tiếng Việt.

Từ khoá: Học sâu; Phân loại văn bản; LSTM; CNN.