IMPROVE CNN AND LSTM IN SENTIMENT ANALYSIS FOR VIETNAMESE FROM DATA PREPROCESSING PHASE

Nguyễn Ngọc Duy*, Lưu Ngọc Điệp*

* Information Technology Department - Posts and Telecommunications Institute of Technology

***Abstract*—**The deep learning method has achieved good results in many application fields, such as image processing and computer vision. Recently, this method has also been used in the field of natural language processing and has achieved good results too. In this area, an issue of concern is subjective opinion classification. A subjective opinion is an individual's thinking or judgment about a product or a socio-cultural event or issue. Subjective opinions have received attention from many producers and businesses who are interested in exploiting the opinions of the community and scientists. This paper experiments with the deep learning model convolution neural network (CNN), long short-term memory (LSTM), and the combined model of CNN and LSTM. The training data set comprise reviews of cars in Vietnamese that are pre-processed according to the method of aspect analysis based on an ontology of semantic and sentimental approaches. This data set experiment with CNN, LSTM, and CNN + LSTM models are used to evaluate the effectiveness of the data preprocessing method that was used in this paper. This paper tests the sentiment classification with the English Sentence Collection Stanford Sentiment Treebank (SST) to assess the validity of the test models with the Vietnamese opinion set. The non-neural method, SVM, was also tested to evaluate the effectiveness of the data processing method of the paper.

***Keywords*—**Classification, CNN, Convolution Neural Network, Corpus, Deep Learning, Long Short Term Memory, LSTM, Opinion Mining, Sentiment Analysis, SVM.

I. INTRODUCTION

Everyone has the opportunity to express their thoughts and assessments about a product, an event, or other problems in the world due to the development of the Internet. Such user opinions are increasing in the internet environment. This is a particularly useful resource for individuals and organizations who want to exploit the opinions of the community, as well as individuals who consider or want to buy a product. There is a large demand for the exploitation of this resource. With a substantially large number of opinions, the classification to exploit them using scientific methods represents an indispensable requirement. Many methods have been proposed and tested for the problem of sentiment analysis of each opinion. Machine learning methods are highly popular in this field, such as Support Vector Machine, Naïve Bayes, and Conditional Random Fields [1][2]. The results have been highly promising. These machine learning methods are based on a sentiment dictionary with a predefined weight for each item. An important feature of a sentiment dictionary is that it was built for a specific language and certain topics. Therefore, if a dictionary is not sufficiently good or used for an inappropriate topic, it can affect the quality of the sentiment analysis.

Deep learning has only relatively recently been applied to natural language processing but has demonstrated good results, such as Nal Kalchbrenner et al. in [3] achieving an accuracy of 86.8% and Qiudan Li et al. in [4] producing an accuracy of 89.81%. The Deep learning method differs from other machine learning methods mentioned above in a particularly basic way in that is no requirement for a sentiment dictionary. Building a sentiment dictionary is a difficult task, especially for the less popular languages in the world.

Methods that need a sentiment dictionary are a major obstacle to less popular languages in the world. Therefore, the approach of the deep learning method brings substantial opportunities for less popular languages such as Vietnamese to develop the sentiment analysis field in natural language processing quickly. Vietnamese has not yet had a sentiment dictionary commonly for research. For less popular languages when using deep learning methods, the important focus is testing, which requires substantial time to identify the specific characteristics of these languages. From there, it is possible to develop solutions to exploit deep learning methods for this field in an effective manner.

Studies on deep learning methods for the field of sentiment analysis in Vietnamese, such as [5][6], [7] also obtained good results. However, the number of such studies has yet to reach a sufficient level. The Vietnamese corpus is not yet rich in the topic. This paper develops the method of [5] in terms of enriching the corpus and ontology. In addition, this research conducted additional tests of some deep learning methods in sentence- and aspect-level analysis. From there, some characteristics can be drawn from using the deep learning method for sentiment analysis in Vietnamese as well as a more detailed evaluation of the method proposed in [5].

The rest of the paper is organized as follows: Section 2

Tác giả liên lạc: Nguyễn Ngọc Duy
Email: duynn@ptithcm.edu.vn
Đến tòa soạn: 09/2020, chỉnh sửa: 10/2020, chấp nhận đăng: 10/2020
Một phần kết quả của bài báo được trình tại hội nghị quốc tế ICICT 2020 London.

reviews the related works. Section 3 gives an overview of the deep learning models used in testing and our preprocessing data method. Section 4 discusses the configuration of the deep learning system and the experimental results. Finally, Section 5 presents conclusions.

II. RELATED WORKS

In the form of sentence-level sentiment classification, Nal Kalchbrenner et al. [3] used the dynamic convolution neural networks model to model sentences by a Dynamic K-Max Pooling operator as a nonlinear sampling function to test sentiment classification using the SST corpus. The accuracy of [3] was 86.8%. Xingyou Wang et al. [8] also tested on the SST opinion set but by using a combination of CNN and RNN. The accuracy of [8] was 89.95% on SST with three sentiment labels. The accuracy of [8] was 51.5% only if Xingyou Wang et al. used SST with five sentiment labels. For Pang and Lee's film rating, 2005, the accuracy of [8] was found to be 82.28%.

In the form of aspect-level sentiment classification, Quidan Li et al. [4] used the convolutional neural network to obtain an accuracy of. The corpus used in the experiment was Chinese blog sites. Meanwhile, Dhanush D. et al. [9] obtained an accuracy of only 76.1% when using CNN to classify the sentiment of restaurant reviews.

III. APPROACH METHOD

As mentioned in the previous sections, this research will use some models of deep learning methods to implement sentiment classification. In the following sections, the paper presents an overview of the LSTM and CNN models, as well as the combined model of CNN and LSTM for sentiment classification testing.

A. The models for experiments:

1) Convolutional Neural Network: A convolution neural network (CNN) is first used in the field of digital signal processing. Based on the principle of information conversion, scientists have applied this technique to digital photo and video processing. In the CNN model as Fig. 1. The layers are linked together through a convolution mechanism. The next layer is the cumulative result from the previous layer. As a result, we have local connections. Hence, each neural in the next layer is generated from the filters imposed on a local data area of the previous neural layer. As a result, we have local connections.

Each CNN contains a word-embedding layer, a convolutional layer, a pooling layer, and a fully connected layer.

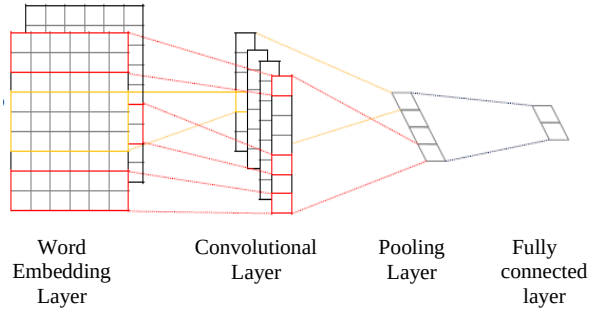

Fig. 1. Basic Convolutional Neural Network model [10].

Word Embedding Layer. This class includes matrices of size *n x k*, representing sentences with *n* words, each representing a *k*-dimensional vector.

This class encodes every word in the selected sentence into a word vector. Let l ∈ R be the sentence length, $|D| \in R$ is the vocabulary size and $W^{(l)} \in R^{kx|D|}$ is the matrix embedded vectors from *k* dimensions. The *i* word in the sentence is transformed into a *k* dimensional vector w_i using equation (1):

$$w_i = W^{(l)} x_i \quad (1)$$

where x_i is a one-hot vector representation for the *i* word one-hot vector.

We can use the tool word2vec or Glove to create the matrix for this layer.

Convolution Layer. This class uses convolution to process data by sliding a fixed size slide (also called a kernel) on the input data matrix to obtain a refined result.

Pooling Layer. This layer is responsible for summarizing the result vector of the convolution layer and retaining the most important vectors.

Fully Connected Layer. In a convolution neural network, there are one or more fully connected layers after the convolution layer. This class is simply a traditional neural network that uses the remaining vectors in the upper layers as input to produce the final result through training.

2) Long Short Term Memory: Long Short-Term Memory (LSTM) has four interaction layers and two status signals: hidden state and cell state, as shown in Fig. 2.

At the time *t*, the LSTM decides that information poured into the cell state based on the sigmoid function or the σ, floor, forget gate. This function takes h_{t-1} from the previous hidden layer and input signal x_t in the present to create a number in [0, 1] as in formula (2). Whether the new information is saved to the cell state or not is based on the calculation at the input port with the sigmoid function as formula (3). A vector of new candidate values $\tilde{C}_t$ is created through the *tanh* layer using formula (4). The cell state at time t is calculated by (5) based on the previous cell state C_{t-1}, candidate $\tilde{C}_t$ và f_t function. The f_t function can control the slope that passes through it and explicitly deletes or updates memory. The LSTM network decides the output based on the cell state. Through the sigmoid function, the value of the cell state is calculated by formula (6) and reached at the output gate. This value is via the *tanh* function and multiplied by the output of the sigmoid port to obtain the h_t value via formula (7).

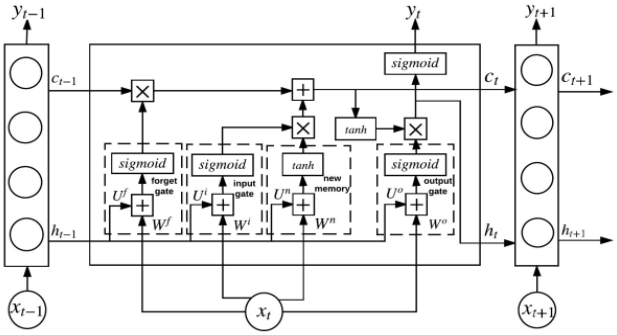

Fig. 2. Long Short Term Memory network [11]

$$ft = \sigma(Wfxt + Ufht\text{-}1 \quad (2)$$

$$i_t = \sigma(W^i x_t + U^i h_{t-1}) \quad (3)$$

$$\tilde{C}_t = \tanh(W^n x_t + U^n h_{t-1}) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W^0 x_t + U^0 h_{t-1}) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

3) CNN – LSTM: Convolution layer of CNN creates a feature map vector. The number of feature vectors is equal to the number of filters used during the convolution. In the pooling layer, the best feature map values from each class will be chosen to obtain the most important feature of the comment. The feature vectors over a fully connected CNN network generate a set of parameters at the output of CNN. The LSTM Ministry uses output parameters of CNN to carry out the process of classifying comments. This combined model is shown in Fig. 3.

4) Support Vector Machine [12]: SVM is a supervised machine learning algorithm introduced by Vladimir N. Vapnik in 1995. The basic idea of SVM is to find a hyperplane to separate data. This hyperplane will divide the space into different domains that contain a type of data. In a two-dimensional space, this hyperplane is a line that separates a plane into two parts that each layer is on the sides. The distance from the points on the hyperplane (called margins) is as far as possible so that the error of classification is minimal.

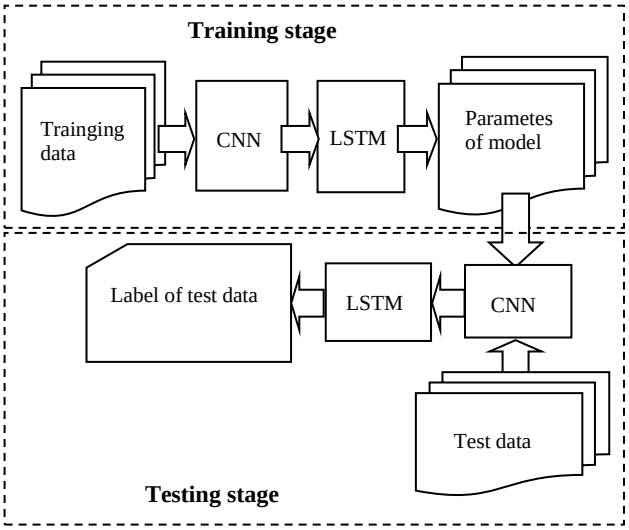

Fig. 3. CNN-LSTM model

Consider a data set D in two-dimensional space that can classify into two classes. Represent data set D as formula (8).

$$D = \{(x_i, y_i) \mid x_i \in R^m, y_i \in \{-1,1\}\}_{i=1}^n \quad (8)$$

where y_i has a value of 1 or -1 defines the class of point x_i. Each x_i is a real vector with m dimensions. We need to find the hyperplane with the largest margin separating the points $y_i = 1$ and the points with $y_i = -1$.

To find the optimal hyperplane, we need to solve the optimization problem. The first we find out maximize $W(\alpha)$ in equation (9).

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j), \alpha_i \le 0 \quad (9)$$

where N is the number of samples in the training set; x_i is one of the training vectors; α_i is the weight of the training sample x_i; K is a kernel function which used to measure the similarity between two samples. If $\alpha_i > 0$, x_i is called a support vector.

For a sample x_j of unknown class, the classification corresponds to solving (10).

$$w = \sum_{i=1}^{r} \alpha_i y_i x_i \quad (10)$$

If we have accepted an error in the training set (there are noise samples) so we have to determine the minimum of the expression (3).

$$\| w \|^2 + C \sum_{i=1}^{N} \xi_i \, , 0 \le \alpha_i \le C \quad (11)$$

where ξ is a slack variable and allows training patterns that exist in the region between margins; the constant C is used to adjust for the best results when classified by SVM.

B. Corpus:

1) English corpus: Stanford Sentiment Treebank is a corpus of film reviews in English [13]. Opinions are categorized using five labels: very positive, positive, neutral, negative and very negative. To have similarities with the Vietnamese corpus, this paper will carry out an empirical classification of sentiments using three labels; the very positive and the positive will be combined into one positive label; very negative and negative classes into one negative label.

2) Vietnamese corpus: The Vietnamese corpus is a collection of documents that record the review of car objects (Car Opinions in Vietnamese - COV). This topic has attracted the attention of many people in large countries such as the United States and Germany where car ownership is particularly popular, and Vietnam where such ownership is popular in some large cities.

Opinions were collected from the online newspapers that record readers' opinions in the auto categories, auto forums, and websites of businesses that sell this item. This corpus was built and processed according to the method presented in [5]. Collected opinions that have been processed by this method comprise a so-called standardized opinion set (COVn). Opinions that have not been processed by this method are referred to as raw comments (COVr). The pretreatment method according to [5] can be summarized as follows:

a) Creating a set of aspects for the objects: The aspects of cars are technical characteristics that often used by manufacturers when introducing products, or customers are interested in and assessing products. Each aspect has an official name that is commonly used by manufacturers and a semantic equivalent. They are the common name that

users use when expressing their opinions in the internet environment.

The aspect entities are collected by two methods:

Statistical methods. This method produces statistics for the number of nouns that appear in the corpus, from which the automotive aspect nouns are selected.

The word2vec tool. This method is based on the characteristics of the correlation of words/phrases in the corpus to define aspect and sentiment items further.

b) Creating the structure of a semantic and sentiment vocabulary hierarchical tree: Both aspect and sentiment entities are divided into three tiers in the SSVHT ontology (Semantic and Sentiment Vocabulary Hierarchical Tree), as shown in Fig. 2. Aspect set refers to words/phrases used by the manufacturer or user to refer to the components or specifications of a car. For example: engine, steering wheel, dashboard.

c) Labeling sentences with the semantic and sentiment on aspects of object: The labeling of semantics and sentiments of the sentence according to the aspect of the car is done manually. The labeling process is divided into two stages. The first stage carries out the labeling according to the aspect of the sentence. The next stage is the sentimental tagging. The process of labeling will remove sentences without Vietnamese accents, without sentiment, or not the same subject. Semantic labels are nouns or noun phrases belonging to the aspect set of cars. The sentimental labels include positive, neutral, and negative.

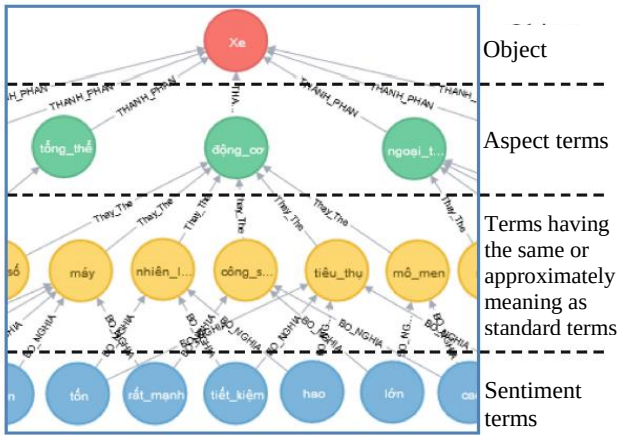

Fig. 4. Semantic and sentiment vocabulary hierarchical tree [5].

d) Standardizing sentiment aspects of the object: Sentiment sentences in the corpus will be re-formulated into the simple structured sentence*:*

$$N + A \quad (11)$$

$$N + P + A \quad (12)$$

$$N + A + P \quad (13)$$

Where:

N: noun representing an aspect.

R: adverb or adverbs

A: sentiment complements (adjectives or verbs).

The features of the corpus after the above processing are shown in Table I.

Table I. The Organization of Vietnamese Corpus

Feature of corpus	Quatity
Sample of cars	102
Opinios	3,214
Simple aspect-based sentences	6,517
Testing sentences	1,130
Sentiment labels	3 (positive, neutral, negative)
Sentences labeled with positive polarity	2,134
Sentences labeled with negative polarity	2,055
Sentences labeled with neutral polarity	2,328

C. *Experiment configuration*

1) System configuration: Table II presents the components of the system used in the experiment of this paper.

Table II. System Specifications

Component of System	Version name
Operating System	Ubuntu 20.04 LTS
Programminng Language	Python
Framework Deep Learning	Theano - Keras

2) Specification of models:

CNN. Activation function is sigmoid ($1/(1+e^{-x})$), embedding word size is 300, number of filters is 300, filter size is 3, dropout is 0.5.

LSTM. Activation function is sigmoid ($1/(1+e^{-x})$), dropout is 0.2, embedding word size is tested at 300, number of filters tested at 100 and 200.

CNN+LSTM. The configuration parameters of CNN and LSTM when combined are the same as when operating independently.

SVM. Build the feature vector of the sentence by the tf-idf method [14]. The kernel function is rbf (Radial Basis Function).

IV. EXPERIMENT

This paper attempts to change the number of filters in the LSTM model. The experiment results with the SST, COVr, and COVn corpus are shown in Figs. 5, 6, and 7.

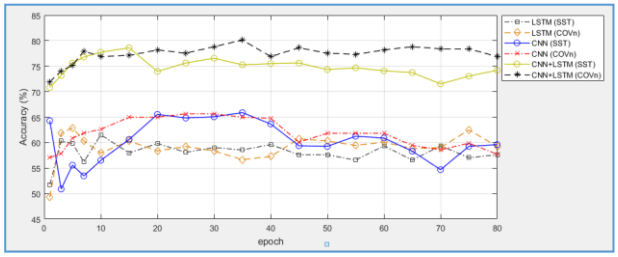

Fig. 5. Accuracy of experimenting on SST and COVn with LSTM has 200 filters.

Figs. 5, 6, 7 and Table III show the results of sentiment classification by the CNN, LSTM, and CNN+LSTM model on corpus SST, COVr, and COVn.

The best values were obtained during the testing of three sets of comments by the LSTM, CNN, and CNN+LSTM model according to the number of iterations shown in Table III. The LSTM has 200 filters and CNN with 300 filters.

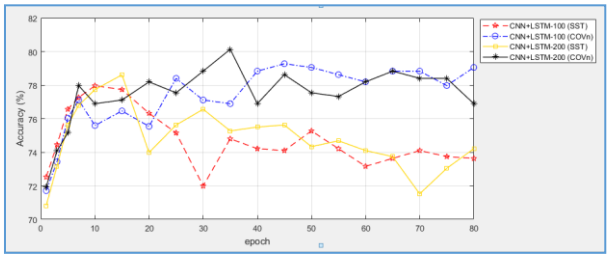

Fig. 6. Accuracy of experiments on SST and COVn with LSTM has 100 and 200 filters.

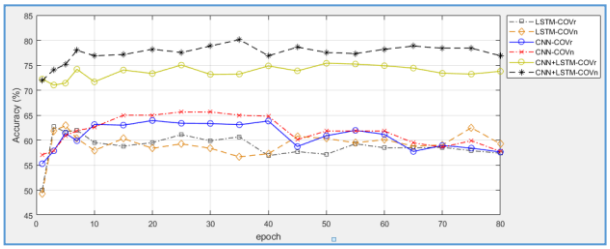

Fig. 7. Accuracy of experiment on COVr and COVn with LSTM has 200 filters.

Table III. Best Accuracy Results of Each Model.

Model	Accuracy (%)	epoch
LSTM (100 filter – COVr)	61.65	5
LSTM (100 filter – COVn)	63.65	5
LSTM (100 filter – SST)	61.12	3
LSTM (200 filter – COVr)	62.66	3
LSTM (200 filter – COVn)	62.91	5
LSTM (200 filter – SST)	61.55	10
CNN (300 filter – COVr)	66.77	25
CNN (300 filter – COVn)	70.64	25, 30
CNN (300 filter – SST)	65.89	20
CNN + LSTM (100 filter – COVr)	75.03	25
CNN + LSTM (100 filter – COVn)	78.83	40
CNN + LSTM (100 filter – SST)	70.86	10
CNN + LSTM (200 filter – COVr)	75.39	50
CNN + LSTM (200 filter – COVn)	**80.13**	35
CNN + LSTM (200 filter – SST)	78.61	15
SVM (COVr)	73.36	-
SVM (COVn)	73.43	-

Some results of classifying comments are presented in Table IV. Sentences are normalized for aspect-based analysis before being sentiment analyzed.

From the experiment results shown in Figs. 5, 6, 7 and Table III, the following statements can be made:

- Both CNN and LSTM models operate independently, obtaining a fairly low accuracy when they analyze aspect-level sentiment. The CNN and LSTM models operating independently demonstrated an accuracy that fluctuated significantly when the number of iterations (epoch) was low (less than 50 times). Within 80 iterations, the best results showed low iterations for both SST and COV opinion sets (from 5 to 10 for LSTM and 25 to 30 times for CNN). By increasing the number of iterations and filters, the result was more stable and reliable.

- The accuracy of the combined CNN and LSTM model is better than the accuracy of each model when they run independently, as shown in Table III. This accuracy improved by approximately 10% compared to when these models were running separately.

- The LSTM model was the least effective of the models tested. This model ran quite slowly and had the lowest accuracy for all three test sets tested.

- The aspect-level data preprocessing method for comments based on SSVHT ontology improves accuracy in both CNN, LSTM, and CNN + LSTM models. In particular, the CNN + LSTM model has an accuracy difference of up to 4% when it was trained by the preprocessed dataset compared to the raw dataset.

- The SVM method achieved accuracy not high. The difference in accuracy when testing on two datasets is also not high. The SVM method is not based on the neural model as deep learning methods. Therefore, it can be seen that the data preprocessing method of the paper is consistent with the deeper learning models

Although the results obtained on SST and COV tests differed with respect to accuracy, there were many similarities in the results achieved in each model as well as between models. Hence, CNN and LSTM in particular, and deep learning in general, do not have any language barriers as well as the subject of sentence-level and aspect-level sentence sentiment classification.

Table IV. RESULTS OF CLASSIFICATION AND DISCUSSION OF THE MODELS.

Sentence		Type	Label	CNN	LSTM	CNN+LSTM
1	*Xèng* hạt dẻ cho một em hạng B	Raw	Positive	Neutral	Neutral	Neutral
	Giá bán hạt dẻ cho một em hạng B	Standardized	**Positive**	**Positive**	**Positive**	**Positive**
2	Các bác nghĩ sao chứ *trông* em nó ẻo lả quá ạ	Raw	Negative	Neutral	Neutral	Neutral
	Các bác nghĩ sao chứ *tổng thể* em nó ẻo lả quá ạ	Standardized	**Negative**	**Negative**	**Negative**	**Negative**
3	Công nhận *khoang lái* em này nhìn khá chất	Raw	Positive	Positive	Positive	Positive
	Công nhận *nội thất* em này nhìn khá chất	Standardized	**Positive**	**Positive**	**Positive**	**Positive**
4	*Lái* qua những đoạn xóc em cảm thấy nó có vẻ ọp ẹp lắm	Raw	Negative	Negative	Negative	Negative
	Vận hành qua những đoạn xóc em cảm thấy nó có vẻ ọp ẹp lắm	Standardized	**Negative**	**Negative**	**Negative**	**Negative**
5	*Ôm vô lăng* trên cao tốc mới thấy nó chao liệng như thế nào nhé bác	Raw	Negative	Neutral	Neutral	Negative
	Vận hành trên cao tốc mới thấy nó chao liệng như thế nào nhé bác	Standardized	**Negative**	**Negative**	**Negative**	**Negative**

V. CONCLUSION

This paper achieved good results for aspect-level sentiment analysis using the CNN and the LSTM model, particularly the model combining CNN + LSTM by helping the data preprocessing method based on an ontology of semantic and sentimental approaches. This data preprocessing method helps the learning process of CNN and LSTM models, and combined models can be performed more rapidly if the CNN and LSTM are separated. This data preprocessing method enables the CNN+LSTM model to achieve good results with a small corpus. The importance of this method is in understanding the subject of the corpus. The data preprocessing method was tested in the aspect-level sentiment classification. Next, the authors will test document-level sentiment classification as well as improve the method so that we can obtain good results for that problem.

The test results of this paper on the SST corpus in English and COV corpus in Vietnamese demonstrate that the deep learning method does not meet the limitations for different languages. This feature of the deep learning method is of high significance to the less popular languages in the world, including Vietnamese. It will help these languages develop the sentiment analysis field in natural language processing more rapidly.

Acknowledgments: This article received some valuable advice from Prof. Dr. Phan Thi Tuoi. The authors are grateful to Professor.

REFERENCES

[1] Balazs, Jorge A., and Juan D. Velásquez, "Opinion mining and information fusion: a survey", Information Fusion Vol 27, 2016, pp 95-110.

[2] MORAES, Rodrigo; VALIATI, JoãO Francisco; NETO, Wilson P. GaviãO, "Document-level sentiment classification: an empirical comparison between SVM and ANN", Expert Systems with Applications, Vol 40, No. 2, 2013, pp 621-633.

[3] Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom, "A convolutional neural network for modelling sentences". arXiv preprint arXiv:1404.2188, 2014.

[4] Qiudan Li, Zhipeng Jin, Can Wang, Daniel Dajun Zeng, "Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems", Knowledge-Based Systems, Vol 107, 2016, pp 289-300.

[5] Duy Nguyen Ngoc, Tuoi Phan Thi and Phuc Do, "A data preprocessing method to classify and summarize aspect-based opinions using deep learning", Asian Conference on Intelligent Information and Database Systems. Springer, 2019. pp. 115-127.

[6] Vo, Q. H., Nguyen, H. T., Le, B., & Nguyen, M. L., "Multi-channel LSTM-CNN model for Vietnamese sentiment analysis", In: 2017 9th international conference on knowledge and systems engineering (KSE). IEEE, 2017, pp 24-29.

[7] PHAM, Thai-Hoang; LE-HONG, Phuong, "The importance of automatic syntactic features in Vietnamese named entity recognition", arXiv preprint arXiv:1705.10610, 2017.

[8] Xingyou Wang, Weijie Jiang, Zhiyong Luo, "Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts". In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016. pp 2428-2437.

[9] Dhanush, D., Thakur, A. K., & Diwakar, N. P, "Aspect-based sentiment summarization with deep neural networks", International Journal of Engineering Research & Technology, Vol 5, Issue 5, 2016, pp 371-375.

[10] Yoon Kim, "Convolutional neural networks for sentence classification", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp 1746-1751.

[11] Lei Zhang, Suai Wang, and Bing Liu, "Deep learning for sentiment analysis: a survey", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol 8, Issue 4, 2018, pp e1253.

[12] VAPNIK, Vladimir; VAPNIK, Vlamimir. Statistical learning theory Wiley. *New York*, 1998, 1: 624

[13] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. "Recursive deep models for semantic compositionality over a sentiment treebank". In Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp 1631-1642.

[14] DIERK, S. F. The SMART retrieval system: Experiments in automatic document processing—Gerard Salton, Ed.(Englewood Cliffs, NJ: Prentice-Hall, 1971, 556 pp., $15.00). *IEEE Transactions on Professional Communication*, 1972, 1: 17-17.

TĂNG CƯỜNG HIỆU NĂNG CHO LSTM AND CNN TRONG PHÂN TÍCH CẢM XÚC TIẾNG VIỆT TỪ GIAI ĐOẠN TIỀN XỬ LÝ DỮ LIỆU

Tóm tắt—Phương pháp học sâu đã đạt được kết quả tốt trong nhiều lĩnh vực ứng dụng, chẳng hạn như xử lý hình ảnh và thị giác máy tính. Gần đây, phương pháp này cũng đã được sử dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên và cũng đạt được kết quả tốt. Trong lĩnh vực này, một vấn đề cần quan tâm là phân loại ý kiến chủ quan. Ý kiến chủ quan là suy nghĩ hoặc nhận định của cá nhân về sản phẩm, sự kiện hoặc vấn đề văn hóa xã hội. Ý kiến chủ quan đã nhận được sự quan tâm của nhiều nhà sản xuất, kinh doanh quan tâm, khai thác ý kiến của cộng đồng và các nhà khoa học. Bài báo này thử nghiệm với mạng nơ-ron tích chập của mô hình học sâu (CNN), bộ nhớ ngắn hạn dài (LSTM) và mô hình kết hợp của CNN và LSTM. Bộ dữ liệu đào tạo bao gồm các bài đánh giá về ô tô bằng tiếng Việt được xử lý trước theo phương pháp phân tích khía cạnh dựa trên bản thể luận của các phương pháp tiếp cận ngữ nghĩa và cảm tính. Thử nghiệm tập dữ liệu này với các mô hình CNN, LSTM và CNN + LSTM được sử dụng để đánh giá hiệu quả của phương pháp tiền xử lý dữ liệu đã được sử dụng trong bài báo này. Bài báo này kiểm tra sự phân loại ý kiến với Ngân hàng Bộ sưu tập câu tiếng Anh Stanford Sentiment Tree (SST) để đánh giá tính hợp lệ của các mô hình thử nghiệm với tập ý kiến tiếng Việt. Phương pháp phi thần kinh, SVM, cũng đã được thử nghiệm để đánh giá hiệu quả của phương pháp xử lý dữ liệu của bài báo

Từ khóa—CNN, học sâu, khai phá ý kiến, kho ngữ liệu, LSTM, mạng neural tích chập, phân tích cảm xúc, SVM.

Nguyen Ngoc Duy is currently a lecturer of the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam, campus Ho Chi Minh City. I received M.Sc. in Computer Science in the Ho Chi Minh City University of Technology, Vietnam (HCMUT) in 2005, and became Ph.D. Candidate at HCMUT since 2016. My research interests include machine learning, data mining, and natural language processing.

Email: duynn@ptithcm.edu.vn

Luu Ngoc Diep is currently a lecturer of the Faculty of Information Technology at Posts and Telecommunications Institute of Technology in Vietnam, campus Ho Chi Minh City. I received M.Sc. Electronics and Telecommunications in the Ho Chi Minh City University of Technology, Vietnam (HCMUT) in 2003.

Email: luungocdiep@ptithcm.edu.vn