VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: Nguyen Thi Cham 2. Sex: Female

3. Date of birth: June 29th, 1982 4. Place of birth: Hanoi

5. Admission decision number: 654/QĐ-CTSV Dated: September 5th, 2016

6. Changes in academic process:

7. Official thesis title: **Developing knowledge distillation techniques in lifelong learning for the text data domains**

8. Major: Information System 9. Code: 9480104.01

10. Supervisors: Assoc. Prof. Ha Quang Thuy

11. Summary of the **new findings** of the thesis:

The four main findings of the thesis are as follows:

Firstly, the thesis proposed the close domain lifelong topic model algorithm CD-AMC based on the AMC lifelong topic model with the must-link and cannot-link distilling knowledge solution from the close past domains instead of from all past domains. The thesis also proposed a general framework for applying close domain lifelong topic model to text analysis tasks. At the same time, two ways of determining the domain close to the current data domain: based on the predicate level - top word level - topic level and based on the past text classifier. The study deployed and applied the close-domain lifelong topic model to the Vietnamese multi-label classification task and the English sentiment classification task.

Secondly, one-sample statistical test on 20 observations according to the *t-distribution* for the hypothetical but unknown population mean showed that CD-AMC model actually has higher efficiency than AMC, about 0.27%.

Thirdly, the thesis proposed the targeted close domain lifelong topic model named TCD-AMC which combines the CD-AMC model with the targeted topic model TTM, both taking advantage of the useful past knowledge from the close domains and more focused on each aspect specified through the target keywords. The thesis also deployed and applied the TCD-AMC model to the Vietnamese text multi-label classification task base on deep learning. Experimental results on six different setting methods showed that the TCD-AMC model improved performance compared to AMC model and other related topic models on the same dataset.

Finally, the thesis proposed the lifelong machine learning model BiLSTM-KD-NER that refined deep learning model knowledge distillation (past BiLSTM model parameters and past NER classification model) for the Vietnamese biomedical named entity recognition task and conducted experiments (including building experimental datasets) to evaluate the BiLSTM-KD-NER model. The experimental results on different scenarios showed that the efficiency of the proposed model took advantage of the knowledge distilled from the close domains to build a lifelong deep neural network in order to solve the limitation of catastrophic forgetting.

Four scientific articles are published in Scopus and DBLP publications.

12. Practical applicability, if any:

13. Further research directions:

In the next time, the PhD student will focus on studies which may solve some of the remaining limitations of the thesis. In particular, the author will focus on conducting these following studies:

- Firstly, modifying the BiLSTM-KD-NER model to test the model on 20 datasets of 20 product reviews from Amazon.com. In addition, the option to combine the BiLSTM-KD-NER model and the DC-AMC lifelong topic model on these 20 datasets is also considered.
- Secondly, exploiting modern tools to estimate Gibbs sampling parameter to improve LDA model to build topic model A_{N+1} for current domain dataset D_{N+1} with little data. Deeply analyzing the lifelong neural topic model (LNTM) to improve the close domain measure, and at the same time, considering the implementation of DocNADE in order to improve the topics in the topic model for the current task;
- Thirdly, improving must-link knowledge mining of three or more words or must-link knowledge for narrow domain topics.
- Finally, conducting research on methods in problem solving to solve the problem of catastrophic forgetting from neural network learning.

14. Thesis-related publications:

[1] Quang-Thuy Ha, Thi-Ngan Pham, Van-Quang Nguyen, Thi-Cham Nguyen, Thi-Hong Vuong, Minh-Tuoi Tran, Tri-Thanh Nguyen. *A New Lifelong Topic Modeling Method and Its Application to Vietnamese Text Multi-label*

Classification. ACIIDS 2018, pp. 200-210 (**Scopus**, **DBLP**) (2 references from foreign authors).

[2] Thi-Cham Nguyen, Thi-Ngan Pham, Minh-Chau Nguyen, Tri-Thanh Nguyen, Quang-Thuy Ha. *A Lifelong Sentiment Classification Framework Based on a Close Domain Lifelong Topic Modeling Method*. ACIIDS 2020, pp. 575-585 (**Scopus**, **DBLP**) (2 references from foreign authors).

[3] Thi-Cham Nguyen, Thi-Ngan Pham, Hoang-Quynh Le, Tri-Thanh Nguyen, Hong-Nhung Bui, Quang-Thuy Ha. *A Targeted Topic Model based Multi-Label Deep Learning Classification Framework for Aspect-based Opinion Mining*. IEEExplore KSE 2020, pp. 165-170 (**Scopus**, **DBLP**) (1 reference from foreign authors).

[4] Thi-Cham Nguyen, Hoang-Quynh Le, Duy-Cat Can, Quang-Thuy Ha. *Models Distillation with Lifelong Deep Learning for Vietnamese Biomedical Named Entity Recognition*. KSE 2021, pp. 1-6 (**Scopus**, **DBLP**).