INFORMATION OF THE DOCTORAL THESIS

- Title of Thesis: **Machine learning methods for automated information extraction from text**
- Speciality: **Information system**
- Code of speciality: **9.48.01.04**
- Ph.D. candidate: **Nguyễn Thị Thanh Thủy**
- Scientific Supervisors:
    1. Prof. Dr. Từ Minh Phương
    2. Assoc. Prof. Dr. Ngô Xuân Bách
- Academic Institution: **Posts and Telecommunications Institute of Technology**

THE SCIENTIFIC CONTRIBUTIONS:

1. Propose an enhanced solution for Vietnamese aspect extraction and opinion classification by leveraging labeled data sources from other languages. The identification of aspects and sentiment classification is carried out on a sentence-level basis instead of using a whole review, making it more realistic and applicable across various real-world scenarios. The proposed method is quite general and flexible as it is language-independent and agnostic to specific machine learning algorithms. This approach helps address the challenges arising from the lack of training data resources in low-resource languages for this problem (such as Vietnamese).
2. Propose methods to extract information from Vietnamese legal documents using both traditional machine learning and deep learning approaches. The extracted information includes reference entity and the relation between legal entities within the text. In addition to employing traditional machine learning methods, this study also incorporates deep learning techniques, harnessing the advantages of deep learning models while leveraging manually designed features (by traditional machine learning methods).
3. Propose a joint entity and relation extraction method for Vietnamese legal documents using a deep learning-based model. The joint extraction model employs a Transformer-based encoder-decoder architecture with a non-autoregressive decoding mechanism to simultaneously extract both references and relations from legal documents. (This differs from the second proposal, which performs sequential extraction of these information.)

ON PRACTICAL APPLICABILITY AND FURTHER STUDIES:

1. The results of the first proposed research show that by incorporating additional labeled data translated from other languages (such as English), the proposed method improved the effectiveness in both aspect extraction and sentiment classification tasks for Vietnamese. With the obtained results, future research directions can focus on advanced methods that leverage available data from other languages to improve NLP tasks for low-

resource languages (such as Vietnamese). For instance, this can involve: (1) Leveraging multiple datasets of the same NLP task from different languages to deal with the task in Vietnamese language; (2) Leveraging multiple datasets of different NLP tasks from another language to solve those tasks in Vietnamese language (i.e. multi-task learning).

2. The results of the second and third proposed research demonstrate that the proposed models can accurately extract individual as well as simultaneous reference entities and relations between entities in Vietnamese legal documents. This is particularly notable for achieving superior performance compared to other baseline models in complex legal sentences containing multiple entity references. Several potential research directions can be pursued: (1) The results of extracted references and relations from legal documents can be useful for other legal text processing tasks such as legal information retrieval, legal text summarization, and question answering in the legal domain; (2) Building legal text processing applications that help people in reading, understanding, and retrieving necessary information from legal documents. These things make a lot of sense in practice.

Confirmation of Scientific Supervisors **Ph.D. candidate**

Prof. Dr. Từ Minh Phương **Nguyễn Thị Thanh Thủy**

Assoc. Prof. Dr. Ngô Xuân Bách