INFORMATION OF THE DOCTORAL THESIS

Thesis title: *"Research on methods for detecting Vietnamese spam messages"*

Speciality: Information System

Code: 9.48.01.04

PhD. Candidate: Vu Minh Tuan

Scientific supervisors:

1. **Assoc. Prof. Tran Quang Anh, PhD**
2. **PhD. Nguyen Xuan Thang**

Training institution: Posts and Telecommunications Institute of Technology

NEW FINDINGS OF THE THESIS

1. The dataset of spam and regular messages was developed for research purposes by the thesis. The dataset consists of 69,192 Vietnamese messages, both with and without diacritics, collected from various sources such as organizations, spam traps, and volunteers. This dataset plays a crucial role in analyzing the characteristics of Vietnamese spam messages, as well as in testing and evaluating spam detection models in the thesis.
2. The dependency of the effectiveness of spam detection models on the length of the message content was analyzed by the thesis. Based on this analysis, a Vietnamese spam detection model is proposed that remains stable as message length varies. Compared to models using statistical rules and traditional machine learning algorithms, the model employing deep learning—specifically CNN networks—demonstrates superior spam detection capabilities, particularly when the message content is limited. The difference in model effectiveness becomes evident with variations in the length of the test messages.
3. Two approaches to address the problem of spam detection considering the variability of the Vietnamese language were proposed by the thesis. The first approach involves converting the input data from multiform Vietnamese (with and without diacritics) to a uniform form (either with or without diacritics). The second approach retains the input data in its multiform Vietnamese state.

APPLICATIONS, PRACTICAL APPLICABILITY AND MATTES NEED FURTHER STUDIES

The research findings of the thesis serve as a valuable reference for businesses related to telecommunications and security to enhance their ability to control and prevent spam messages and unauthorized advertising. Based on several results achieved in addressing the problem, the thesis can be further developed and expanded in the following directions:

(1) Improving detection performance: Enhancing the model architecture, increasing the ability to classify and detect spam messages.

(2) Expanding the dataset: Expanding and improving the training dataset to ensure diversity and representation of various types of Vietnamese spam messages.

(3) Handling complex linguistic features: Vietnamese has many complex linguistic features such as word variants, tonal marks, and rich contexts. Integrating natural language processing techniques and Vietnamese linguistic characteristics into the model to improve the understanding and classification of different types of spam messages.

(4) Applying in other fields: The Vietnamese spam detection model can be applied in various fields such as advertising messages or news, in different applications such as fraud prevention, data management, and content classification.

Confirmation of representative Scientific supervisor | **PhD. Candidate**

Assoc. Prof. Tran Quang Anh, PhD | **Vu Minh Tuan**