THÔNG TIN LUẬN ÁN

Đề tài nghiên cứu: **Phân tích cảm xúc trên cơ sở trị cảm xúc chuyển dịch theo ngữ cảnh cho tiếng Việt**
Chuyên ngành: Khoa học máy tính
Mã số chuyên ngành: 62.48.01.01
Họ và tên NCS: Trần Khải Thiện
Tập thể hướng dẫn: GS.TS. Phan Thị Tươi
Cơ sở đào tạo: Trường Đại học Bách Khoa – ĐHQG TP. HCM

1. TÓM TẮT

Nhiều công trình, công cụ và ứng dụng phân tích cảm xúc đã được phát triển để khai thác các ý kiến trong nội dung do người dùng tạo trên các trang mạng. Tuy nhiên, hiệu năng của các hệ thống này chưa cao do bản thân phân tích cảm xúc là bài toán xử lý ngôn ngữ tự nhiên phức tạp. Các công trình này vẫn chưa hiệu quả trong việc xử lý một số hiện tượng ngôn ngữ, chẳng hạn như các hiện tượng dịch chuyển cảm xúc và văn bản mang ý kiến hỗn hợp.

Luận án khai thác các trường hợp gây hiện tượng dịch chuyển cảm xúc trong văn bản tiếng Việt nhằm thực hiện hai mục tiêu chính: 1) Thứ nhất, xây dựng kho từ vựng cảm xúc cho tiếng Việt phục vụ phân tích cảm xúc mức từ và cụm từ. 2) Thứ hai, tiếp cận phương pháp định hướng ngữ nghĩa kết hợp với các kỹ thuật học máy, mô hình học sâu vào học tổ hợp nhằm xử lý bài toán phân lớp cảm xúc.

Thực nghiệm cho thấy việc quan tâm đến dịch chuyển cảm xúc và việc sử dụng kết hợp nhiều phương pháp là chìa khóa để hệ thống có được kết quả chính xác hơn.

Luận án có 12 bài báo đã công bố, gồm 06 bài đăng trong danh mục tạp chí quốc tế (3 bài thuộc SCIE), 01 bài đăng tạp chí trong nước, và 05 bài trong các đăng kỷ yếu hội nghị khoa học quốc tế.

2. CÁC ĐÓNG GÓP CHÍNH CỦA LUẬN ÁN

Xây dựng từ điển cảm xúc cho từ và cụm từ tiếng Việt.

Thông qua phân tích đặc trưng ngôn ngữ và sự dịch chuyển cảm xúc trong các nhận xét tiếng Việt, luận án đã tiến hành chuyển ngữ sang tiếng Việt các từ cảm xúc tiếng Anh dựa trên từ điển cảm xúc SentiWordnet; sử dụng Hồi quy Logistic và áp dụng tính toán mờ do Zadeh đề xuất để đưa ra mô hình hiệu quả cho việc xác định độ đo cảm xúc của từ và cụm từ tiếng Việt. Luận án điều chỉnh các hàm mờ cho việc tính toán độ đo cảm xúc cụm từ dựa trên cấu trúc cú pháp của cụm từ tiếng Việt để phù hợp với đặc trưng ngôn ngữ tiếng Việt.

Đề xuất mô hình học tổ hợp (*ensemble learning*) hiệu quả với các bộ học thành phần được học trên tập dữ liệu được khai thác nhiều đặc trưng khác nhau của tiếng Việt.

Các đặc trưng khác nhau của tập dữ liệu được xác định bằng phương pháp hướng đến ngữ nghĩa, học máy, và học sâu. Việc lựa chọn mô hình nhúng từ *Word2Vec* và phương pháp học sâu cho bộ học thành phần của mô hình học tổ hợp đã làm cho hiệu năng của mô hình phân lớp cảm xúc được cải thiện. Mô hình đề xuất của luận án có thể áp dụng tốt cho cả ngôn ngữ tiếng Anh.

3. NHỮNG VẤN ĐỀ CÕN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU

Kết quả nghiên cứu của luận án đã giải quyết được một số vấn đề trong việc xử lý bài toán phân tích cảm xúc, tuy nhiên luận án cần thực hiện các nghiên cứu tiếp để cải thiện chất lượng của công trình:

1. Thực hiện nghiên cứu sâu hơn về dịch chuyển cảm xúc, áp dụng vào bài toán phân tích cảm xúc. Mặc dù điều này là một thách thức lớn vì liên quan nhiều đến lĩnh vực ngôn ngữ học. Ví dụ như các câu nhận xét mỉa mia luôn là bài toán hóc búa đối với xử lý ngôn ngữ tự nhiên mặc dù lại hay xuất hiện trong các nhận xét của người dùng. Bên cạnh đó, cần tiếp tục nghiên cứu xử lý triệt để hơn các trường hợp xuất hiện từ phủ định, động từ khiếm khuyết, từ tăng cường-giảm nhẹ, các hiện tượng tương phản, hiện tượng mâu thuẫn (không tương thích) trong câu, trong đoạn văn bản.

2. Xem xét nâng cấp một số công cụ tiền xử lý như bộ phân tích cú pháp văn phạm phục thuộc. Đây là các công cụ có thể gây ảnh hưởng lớn đến độ chính xác của hệ thống.
3. Việc quan tâm xử lý danh từ và cụm danh từ cũng như mở rộng từ điển cảm xúc trong các nghiên cứu tiếp theo cũng là công việc thiết yếu khi mà nguồn dữ liệu cho phân tích cảm xúc tiếng Việt hiện nay còn rất hạn chế.
4. Trọng tâm của luận án là xử lý cho ngôn ngữ tiếng Việt nhưng ý tưởng và các phương pháp hiện thực của mô hình mà luận án đã đề xuất vẫn có thể áp dụng được cho ngôn ngữ khác, như tiếng Anh.

CÁN BỘ HƯỚNG DẪN	NGHIÊN CỨU SINH
GS.TS. PHAN THỊ TƯƠI	TRẦN KHẢI THIỆN

THESIS INFORMATION

Title**: Sentiment Analysis with Contextual Valence Shifters for Vietnamese**

Major: Computer Science

Major Code: 62.48.01.01

PhD student: Tran Khai Thien

Advisors: Prof. Phan Thi Tuoi

University: Ho Chi Minh City University of Technology, VNU-HCMC

1. ABSTRACT

Various sentiment analysis works, tools, and applications have been developed to exploit opinions in user-generated content on social media. However, the performance of these systems is not great because sentiment analysis itself is a complex natural language processing problem. These works are still ineffective in dealing with some linguistic phenomena, such as context valence shifting and mixed opinion text.

The dissertation explores cases of contextual valence shifting in Vietnamese text to accomplish two objectives: 1) to build a sentiment vocabulary database in Vietnamese and 2) to combine the semantic-oriented approach with machine-learning techniques, and deep-learning methodology to handle the sentiment classification challenge.

Experiments show that paying attention to contextual valence shifting and using a combination of various methods are key for the system to yield more accurate results.

The thesis has 12 published articles, among which 06 articles in the list of international journals (3 articles in SCIE), 01 article in domestic journals, and 05 articles in the proceedings of international scientific conferences.

2. MAIN CONTRIBUTIONS

Building a sentiment dictionary for Vietnamese words and phrases.

Through analyzing linguistic features and emotional shifting in Vietnamese comments, the thesis translates into Vietnamese emotional English words based on SentiWordnet emotion dictionary using Logistic Regression and applying fuzzy computation proposed by Zadeh to propose an effective model for determining the emotional metric of Vietnamese words and phrases. The thesis adjusts fuzzy functions for calculating phrase sentiment based on the syntactic structure of Vietnamese phrases to be in harmony with Vietnamese language features.

Proposing an effective ensemble learning model with component learning sets learned on a dataset composed of many different characteristics of Vietnamese.

The different features of the data set are identified by methods oriented towards semantics, machine learning, and deep learning. The choice of the word embedding model Word2Vec and the deep learning method for base-learners of the ensemble learning model improve the effectiveness of the sentiment classification model. The proposed model of the thesis can be well applied to the English language as well.

3. QUESTION ISSUES TO CONTINUE THE RESEARCH

The thesis's research results have deciphered some problems in solving the problem of emotion analysis and yet it needs carrying out further studies to ameliorate the work's quality:

- Conducting more profound research on emotional shifting, then putting emotional analysis problems into application, in spite of this being a big challenge and involving many a field of linguistics.
- Considering upgrading some preprocessor tools such as a dependent grammar and parsing. These tools can greatly affect the accuracy of the system.
- Paying attention to the processing of nouns and noun phrases as well as expanding the emotional dictionary in following studies also constitutes essential work as the data source for Vietnamese emotion analysis is currently very limited.

– The thesis focus on processing for Vietnamese language, but the ideas and practical methods of the model proposed by the thesis can still be applied to other languages, such as English.

ADVISORS	PHD STUDENT
PROF. PHAN THI TUOI	TRAN KHAI THIEN