ỨNG DỤNG KỸ THUẬT SVM (SUPPORT VECTOR MACHINE) VÀO VIỆC PHÂN LOẠI Ý KIẾN ĐÁNH GIÁ ĐỊA ĐIỂM DU LỊCH TỪ MẠNG XÃ HỘI

Nguyễn Huyền Trang[1]

Ngày nhận bài: 17/11/2022; Ngày phản biện thông qua: 14/01/2023; Ngày duyệt đăng: 31/5/2023

TÓM TẮT

SVM (Support Vector Machine) là một phương pháp học có giám sát dựa trên lý thuyết học thống kê. Mục đích của SVM là sử dụng thuật toán học nhằm xây dựng một siêu phẳng làm cực tiểu hóa độ phân lớp sai của một đối tượng dữ liệu mới. SVM cũng là một trong những kỹ thuật được nhiều nhà nghiên cứu quan tâm để sử dụng trong việc phân loại ý kiến về sản phẩm hay dịch vụ dựa trên khai phá dữ liệu. Bài báo này nhằm mục đích tìm hiểu và ứng dụng triển khai thực nghiệm kỹ thuật SVM trong bài toán phân loại ý kiến đánh giá Vịnh Hạ Long với 1.213 ý kiến bằng tiếng Việt thu thập từ mạng xã hội Facebook. Qua 7 lần thực nghiệm kết quả tốt nhất thu được với độ chính xác gần 77%. Bài báo cũng là một gợi ý tốt cho những cá nhân muốn thử nghiệm các phương pháp học có giám sát vào bài toán phân loại ý kiến đánh giá về một địa điểm du lịch bất kỳ.

Từ khóa: *Phân loại ý kiến, SVM, Support Vector Machine, Học có giám sát, Vịnh Hạ Long, Mạng xã hội.*

1. MỞ ĐẦU

Ngày nay, du khách thường có xu hướng tham khảo ý kiến của những người có kinh nghiệm trước khi quyết định đến một địa điểm nào đó. Các mạng xã hội là nguồn tham khảo tuyệt vời khi rất nhiều người dùng có thể để lại bình luận bày tỏ quan điểm cá nhân về một địa điểm du lịch nào đó. Các ý kiến này được thu thập, phân chia thành các loại tích cực hoặc tiêu cực giúp các du khách khác có thể tham khảo hoặc các nhà quản lý lắng nghe ý kiến khách hàng và cải thiện chất lượng dịch vụ.

Phân loại ý kiến khách hàng về sản phẩm hay dịch vụ dựa trên khai phá dữ liệu đã nhận được sự quan tâm của nhiều nhà nghiên cứu như Bing Liu (2011, 2012), nhóm nghiên cứu của Erik Cambria (2013), Kushal Dave và cộng sự (2003). Nhiều kỹ thuật đã được đề xuất, trong đó có SVM, một phương pháp học có giám sát dựa trên lý thuyết học thống kê. Kỹ thuật này đã được áp dụng trong nhiều công trình nghiên cứu, ứng dụng và thu được những kết quả khả quan. Các công trình này chủ yếu áp dụng cho Tiếng Anh.

Bài toán phân loại ý kiến đối với văn bản tiếng Việt cũng đã được nhiều nhà nghiên cứu quan tâm. Có một số công trình đã sử dụng SVM thực hiện phân loại tài liệu Tiếng Việt về âm nhạc, ẩm thực, thời trang, vi tính, kinh doanh... cho kết quả khả quan như Trần Cao Đệ và Phạm Nguyên Khang (2012), Nguyễn Thị Lan Anh (2013). Việc thực hiện phân loại ý kiến đánh giá về một địa điểm du lịch vẫn còn khá mới mẻ nên tôi đã lựa chọn đề tài này cho nghiên cứu của mình, địa điểm du lịch cụ thể là Vịnh Hạ Long. Các ý kiến này được thu thập từ mạng xã hội Facebook.

Có thể mô tả bài toán một cách dễ hiểu như sau:

Input: Tập ý kiến = $\{d_i\}$

Output: d_i là *Tích cực*, *Tiêu cực* hay *Không xác định*?

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Bài toán phân loại ý kiến

Khai phá dữ liệu văn bản hay phát hiện tri thức từ các cơ sở dữ liệu văn bản (*textual databases*) là quá trình phát hiện tri thức từ các cơ sở dữ liệu có cấu trúc. Phân loại ý kiến là một trường hợp cụ thể của bài toán phân loại văn bản. Các "văn bản" được xét phân loại là ý kiến về một vấn đề nhất định nên tương đối ngắn gọn, không quá phức tạp, thể hiện trực tiếp quan điểm người đưa ra ý kiến.

Theo Bing Liu (2011), bài toán phân loại ý kiến được hiểu như sau: "*Cho một tập hợp các văn bản cần đánh giá D, một bộ phân lớp quan điểm sẽ phân lớp mỗi tài liệu d∈D vào một trong hai lớp: Tích cực (positive) hoặc Tiêu cực (negative). Tích cực có nghĩa là tài liệu d thể hiện quan điểm tích cực và ngược lại nếu d thể hiện quan điểm tiêu cực thì d thuộc lớp Tiêu cực"*.

Có 3 cách tiếp cận phổ biến cho bài toán phân loại ý kiến, đó là:

• Phân loại dựa trên các cụm từ chỉ cảm xúc (*Classification Based on Sentiment Phrases*)

• Phân loại bằng cách sử dụng các phương pháp phân loại văn bản (*Classification Using Text Classification Methods*)

• Phân loại bằng cách sử dụng hàm tính điểm

[1]*Trường Đại học Khoa học, Đại học Thái Nguyên;*
Tác giả liên hệ: Nguyễn Huyền Trang; Email: trangnh@tnus.edu.vn.