TẠP CHÍ KHOA HỌC ĐẠI HỌC KHÁNH HÒA
tapchikhoahoc@ukh.edu.vn

PHÂN TÍCH Ý KIẾN PHẢN HỒI CỦA NGƯỜI HỌC DỰA TRÊN PHƯƠNG PHÁP PHÂN LOẠI CẢM XÚC

Hoàng Ngọc Dương

Trường Sĩ quan Không quân

Tóm tắt: *Trong nghiên cứu này chúng tôi xây dựng mô hình phân tích các ý kiến phản người học thông qua việc tự động phân loại và gán nhãn các ý kiến phản hồi. Công việc chính gồm các bước sau: Xây dựng công cụ lấy dữ liệu từ trang thông tin phản hồi, làm sạch dữ liệu, xây dựng mô hình phân lớp dữ liệu dựa trên tập phản hồi người học. Tiến hành phân tích dựa trên bộ từ điển cảm xúc tiếng Việt. Việc phân loại và dự đoán các nhãn được thực hiện dựa trên phương pháp Support Vector Machine (SVM). Thực nghiệm cho kết quả khả quan trên tập phản hồi về một số nội dung Hoạt động đào tạo, Giảng viên/Cán bộ quản lý, Chương trình đào tạo/Giáo trình, Cơ sở vật chất, thiết bị đào tạo và thư viện.*

Từ khóa: *Phân lớp văn bản; Ý kiến người học; Từ điển cảm xúc; Support Vector Machine*

1. Mở đầu

Việc lấy ý kiến, thông tin phản hồi ở các nước phát triển đã có từ lâu, là một hoạt động phổ biến trong quá trình phát triển. Tại các trường người học đóng vai trò quan trọng trong việc đảm bảo chất lượng đào tạo. Hiện nay tại một số trường đại học đều có các kênh để lấy ý kiến phản hồi từ người học về quá trình đào tạo, các hoạt động giảng dạy của giảng viên, cơ sở vật chất nhà trường…

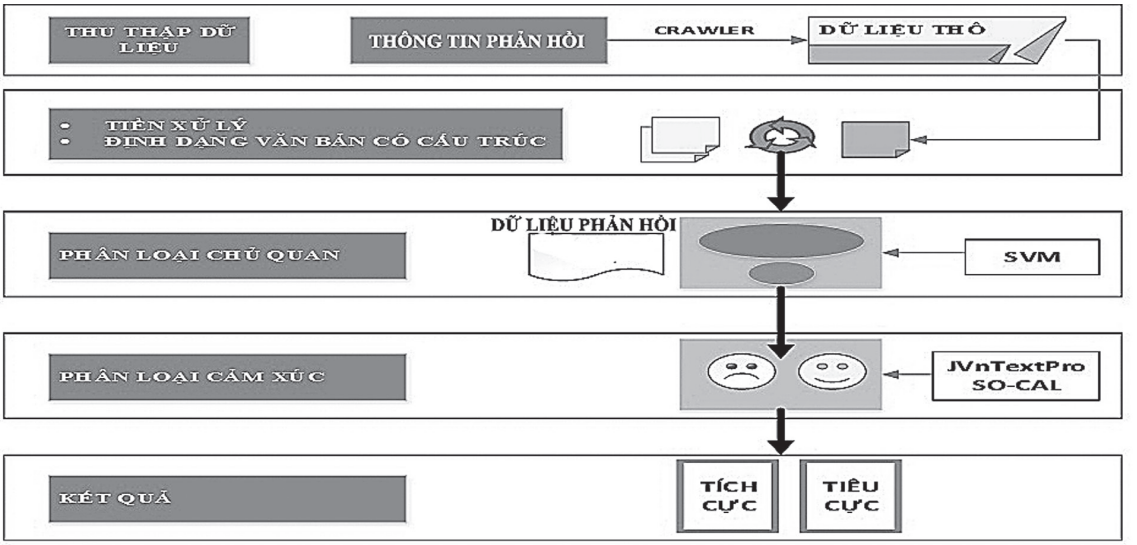
Với mục tiêu tăng cường tinh thần trách nhiệm của sinh viên với quyền lợi, nghĩa vụ học tập, rèn luyện của bản thân, duy trì nâng cao chất lượng đào tạo, trong nhiều năm qua công tác lấy ý kiến phản hồi từ người học về quá trình đào tạo, các hoạt động giảng dạy của giảng viên, cơ sở vật chất nhà trường là nhiệm vụ thường xuyên, liên tục tại cuối mỗi học kỳ, năm học. Việc xử lý các ý kiến phản hồi bằng thủ công gặp rất nhiều khó khăn và tốn nhiều thời gian công sức.

Mục đích của nghiên cứu này, đề xuất phương pháp xử lý các ý kiến phản hồi của người học sử dụng SVM. Nghiên cứu xây dựng bộ từ điển cảm xúc tiếng Việt dựa trên nền tảng bộ cảm xúc tiếng Anh (SO-CAL). Điều chỉnh trọng số phù hợp sau những thử nghiệm nhằm nâng cao hiệu quả chính xác trong việc đánh giá, phân tích, xây dựng mô hình đánh giá dựa trên giải thuật phân lớp SVM đạt hiệu suất tương đối cao.

Các phần tiếp theo được tổ chức như sau: Phần 2 nội dung nghiên cứu: giới thiệu hệ thống phân tích ý kiến phản hồi của người học. Phần 3 là thực nghiệm trình đánh giá kết quả hệ thống, cuối cùng là phần kết luận.

2. Nội dung nghiên cứu

Dựa vào những nghiên cứu lý thuyết và thực tế thông tin phản hồi của người học, chúng tôi xây dựng một mô hình hệ thống thực nghiệm phân tích, đánh giá phản hồi của người học. Trong hình 1 trình bày mô hình thực nghiệm.



Hình 1. Mô hình hệ thống phân tích ý kiến phản hồi

Theo mô hình trên, quy trình phân tích, đánh giá phản hồi người học trải qua các bước như sau: Đầu tiên tiến hành thu thập dữ liệu: thu thập dữ liệu là những nội dung phản hồi tiếng trên trang tin lấy phản hồi. Sau bước này sẽ tiến hành tiền xử lý và định dạng văn bản có cấu trúc. Tiếp đến thực hiện việc định dạng dữ liệu lấy về ban đầu (dạng thô) thành dữ liệu có cấu trúc theo dạng XML, bóc tách các tiêu đề bài viết (title), nội dung bài viết (content), các phản hồi liên quan (comment).

Tiếp theo tiến hành phân loại chủ quan (subjectivity classification): Những phản hồi của người học thông qua giải thuật SVM (đã được huấn luyện) sẽ được phân thành 2 lớp: chủ quan (subjective) và khách quan (objective). Bước cuối cùng thực hiện là phân loại (sentiment classification): Sau khi tập phản hồi được phân làm 2 lớp, thực hiện loại bỏ lớp không có cảm xúc (objective). Áp dụng giải thuật JvnTextPro để tiến hành tách câu, tách từ tiếng Việt và tiến hành so khớp với bộ từ điển để đánh trọng số cho những từ có cảm xúc.

Thu thập dữ liệu: Việc thu thập dữ liệu là một phần công việc rất quan trọng để giải quyết bài toán này. Lượng dữ liệu phải đủ lớn và lấy từ nhiều nguồn khác nhau để nâng cao tính khách quan từ phản hồi của người học.

Thông qua bộ xử lý dữ liệu giao thức HTTP, từ khóa cần tìm thông tin được xử lý và thực hiện dò tìm trong các liên kết (URL). Sau đó, thực hiện lấy dữ liệu thông qua bộ thu thập dữ liệu (Crawler). Dữ liệu lấy về được định dạng để thực hiện cho việc phân tích.

Tiền xử lý, định dạng dữ liệu có cấu trúc: Dữ liệu lấy về ban đầu là dữ liệu dạng thô nên muốn sử dụng cho việc phân tích thì cần xử lý về dạng chuẩn hóa. Dữ liệu được xử lý qua các bước sau: Xử lý tiếng Việt không dấu, xử lý biểu tượng, xử lý "stop words", lấy các thông tin phản hồi. Cuối cùng trong giai đoạn này là định dạng dữ liệu trong file XML.

Dữ liệu huấn luyện: Để nâng cao độ chính xác trong quá trình phân tích, dữ liệu huấn luyện (traning set) đóng vai trò rất quan trọng. Thực hiện xây dựng dữ liệu huấn luyện có chọn lọc theo phương pháp thủ công với 1000 dữ liệu, trong đó có 500 dữ liệu lớp chủ quan (subjective) và 500 dữ liệu lớp khách quan (objective). Dữ liệu huấn luyện tập trung vào các nhóm chính: Hoạt động đào tạo, Giảng viên/Cán bộ quản lý, Chương trình đào tạo/Giáo trình, Cơ sở vật chất, thiết bị đào tạo và thư viện.

Phân loại cảm xúc: Với sự lớn mạnh của truyền thông xã hội trên mạng Internet như forum (diễn đàn), blog và đặc biệt là mạng xã hội (Facebook, Twitter, Instagram,…) thì bài toán phân tích cảm xúc (Sentiment Analysis) đã phát triển nhanh chóng trở thành lĩnh vực nghiên cứu sôi động trong ngành xử lý ngôn ngữ tự nhiên.

Phân loại cảm xúc được thực hiện qua phân tích bình luận, phản hồi…để đánh giá mức độ theo những thang điểm đã được xây dựng trong bộ từ điển cảm xúc tiếng Việt. Từ đó, sẽ có những tổng hợp và phân loại cụ thể.

Bộ từ điển SO-CAL tiếng Việt: Để thông dịch bộ từ điển SO-CAL tiếng Anh chúng tôi

đã sử dụng kết hợp hai bộ từ điển Google Translate và Viettien Dictionary [7].

Sau khi dựa vào hai từ điển trên để thông dịch bộ từ điển SO-CAL tiếng Anh, thu được bộ từ điển SO-CAL tiếng Việt bao gồm 5 bộ từ điển nhỏ: Từ điển danh từ, từ điển động từ, từ điển tính từ, từ điển trạng từ và từ điển từ tăng cường. Một số loại từ trong bộ từ điển SO-CAL tiếng Việt được thể hiện trong bảng 1, 2, 3:

Bảng 1. Một số từ trong bộ từ điển danh từ

Danh từ	Giá trị cảm xúc
hoàn hảo	5
kiệt tác	5
chướng ngại	-2
trò cười	-3
thảm họa	-4

Bảng 2. Một số từ trong bộ từ điển động từ

Động từ	Giá trị cảm xúc
tôn kính	4
thành công	3
sáng tạo	2
xấu hổ	-2
ghét	-4

Bảng 3. Một số từ trong bộ từ điển tính từ

Tính từ	Giá trị cảm xúc
tuyệt vời	5
bổ ích	3
hợp lý	1
bẩn	-3
tai hại	-4

Phương pháp phân loại chủ quan: Phân loại chủ quan là bước đầu tiên cần thiết để phân tích cảm xúc. Trong phần này, công việc cần thực hiện là đánh giá và phân lớp dữ liệu sau khi tiền xử lý dữ liệu thành 02 lớp: lớp chủ quan và lớp khách quan.

Câu có từ hàm chứa cảm xúc: Hiện nay trên thế giới cũng như trong nước, việc phân tích chủ quan chủ yếu dựa vào phương pháp so khớp với bộ từ điển cảm xúc để xác định trọng số cho các từ hàm chứa cảm xúc. Nghiên cứu này lựa chọn phương pháp so khớp từ với bộ từ điển cảm xúc tiếng Việt đã thông dịch.

Phương pháp phân loại cảm xúc: Sau khi xác định được câu có cảm xúc, dựa vào bộ từ điển cảm xúc tiếng Việt và các đặc trưng được rút trích dựa vào những đặc điểm câu văn của tiếng Việt để tính toán giá trị cảm xúc của câu. Dựa vào giá trị này để phân loại câu có cảm xúc thành câu có cảm xúc tích cực hay câu có cảm xúc tiêu cực.

Phương pháp phân lớp Support Vector Machine

Support Vector Machines là một phương pháp học có giám sát để phân lớp dữ liệu. SVM là một công cụ mạnh mẽ cho các bài toán phân lớp phi tuyến tính được Corters và Vapnik giới thiệu vào năm 1995 [6] để giải quyết vấn đề nhận dạng mẫu hai lớp.

Trong những năm gần đây, SVM được biết đến như một hướng tiếp cận phân lớp hiệu quả và đang được áp dụng rộng rãi trong nhiều ứng dụng thực tế. Ưu điểm của SVM là khả năng phân lớp với độ chính xác cao, điều này được đảm bảo bởi các tính chất của siêu phẳng tối ưu và cách sử dụng hàm hạt nhân.

Các bước chính của phương pháp SVM

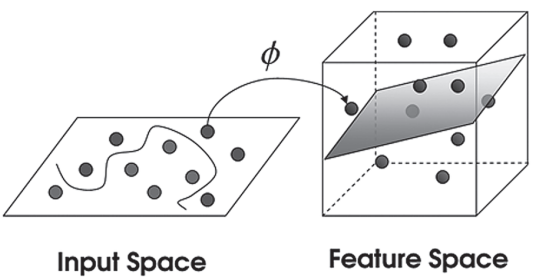
Tiền xử lý dữ liệu: thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả thuộc tính.

Chọn hàm hạt nhân: lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.

Thực hiện kiểm tra để xác định các tham số cho ứng dụng. Điều này cũng quyết định đến tính chính xác của phương pháp này trong quá trình phân lớp.

Sử dụng các tham số cho việc huấn luyện các tập mẫu: trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp bằng cách ánh xạ chúng vào không gian đặc trưng bằng các hàm hạt nhân để giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

Kiểm thử dữ liệu test.



Hình 2. Mô hình Support Vector Machine[3]

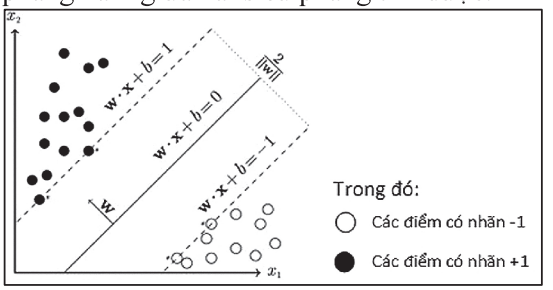
SVM đầu tiên được xây dựng thông qua bài toán phân lớp nhị phân, bài toán phân lớp nhị phân được phát biểu như sau: Cho tập dữ liệu huấn luyện gồm l mẫu $\{(x_1,y_1),...,(x_l,y_l)\}$trong đó đó $x_i \in R^D$ và $y_i \in \{\pm 1\}$, $\forall_i \in \{1,..., l\}$. Trong hình 2, mô tả dữ liệu đầu vào (Input space) chưa được phân lớp và dữ liệu đầu ra (Feature space) là kết quả của quá trình phân lớp.

Tập dữ liệu này được gọi là khả tách nếu tồn tại một hàm tuyến tính $f(x)=w^T.x + b$ để tách tập dữ liệu trên thành hai lớp.

Bài toán phân hai lớp với SVM

Yêu cầu: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai. Cụ thể, với một mẫu dữ liệu x_i thì cần phân vào lớp +1 hay -1.

Để thực hiện bằng phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách chúng là lớn nhất có thể để phân tách hai lớp này ra thành hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được.



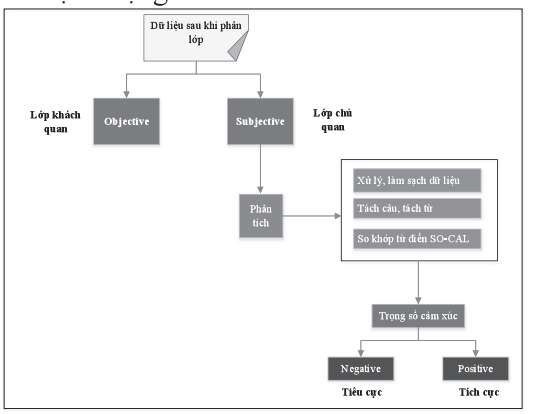
Hình 3. Mô hình phân hai lớp SVM[2]

Trong nghiên cứu này, Phân lớp bằng SVM trải qua hai bước chính như sau: Bước 1: Xây dựng mô hình từ tập dữ liệu huấn luyện. Bước 2: Sử dụng mô hình, kiểm tra tính đúng đắn của mô hình và dùng để phân tích dữ liệu mới.

Mỗi bộ (mẫu) dữ liệu huấn luyện được phân vào một lớp xác định trước. Lớp mỗi bộ (mẫu) dữ liệu huấn luyện được xác định bởi thuộc tính gán nhãn lớp. Trong đó, nhãn (-1) là lớp không có cảm xúc và (+1) là lớp có cảm xúc. Dùng tập các bộ (mẫu) của dữ liệu huấn luyện để xây dựng mô hình huấn luyện. Dựa vào dữ liệu huấn đã phân thành 2 lớp: chủ quan (subjective) và khách quan (objective), tiến hành thực hiện huấn luyện SVM để giúp nâng cao độ chính xác cho những thử nghiệm trên dữ liệu kiểm tra (test set).

Phân lớp chủ quan: Sau bước tiền xử lý và định dạng văn bản có cấu trúc dạng XML, tiến hành sử dụng SVM đã được huấn luyện để áp dụng phân lớp cho tập phản hồi thành 2 lớp chủ quan (subjective) và khách quan (objective).

Phân lớp: Giá trị cảm xúc trong câu phụ thuộc vào loại từ và giá trị cảm xúc của loại từ đó được so khớp với bộ từ điển SO-CAL tiếng Việt. Sau bước phân lớp chủ quan, tập phản hồi đã được phân thành 2 lớp: chủ quan (subjective) và khách quan (objective). Để đánh giá cảm xúc, loại bỏ lớp khách quan (objective) và tiến hành phân tích lớp chủ quan (subjective). Trong phần phân tích này gồm có một số công đoạn như: làm sạch dữ liệu, tách câu, tách từ, so khớp với từ điển để xác định trọng số cảm xúc.



Hình 4. Mô hình phân tích phản hồi người học bằng SVM

Trong quá trình phân tích lớp chủ quan (subjective), tiến hành thực hiện qua các bước sau: Tiền xử lý, tách câu, tách từ tiếng Việt, so khớp từ điển cảm xúc SO-CAL, đánh trọng số dựa vào SO-CAL. Đánh giá phần trăm cảm xúc tích cực (positive) và tiêu cực (negative) dựa vào kết quả trọng số đã được gán cho mỗi phản hồi. Các bước thực hiện:

Bước 1: Tiền xử lý là giai đoạn làm sạch, làm gọn dữ liệu phản hồi như loại bỏ các dấu chấm câu, khoảng trắng, ký tự đặc biệt… Đây cũng là quá trình để chuẩn bị cho việc tách câu, tách từ tiếng Việt và các bước tiếp sau đó.

Bước 2: Tách câu, tách từ. Hiện nay, Có rất nhiều công cụ để thực hiện tách câu, tách từ tiếng Việt. Trong mô hình này, sử dụng bộ thư viện tách từ Jvntextpro được phát triển bởi tác giả Nguyễn Cẩm Tú được cho tại địa chỉ: http://jvntextpro.sourceforge.net/ Đây là bộ thư viện mã nguồn mở trong java.

Bước 3: So khớp từ điển cảm xúc SO-CAL. Sau quá trình tách từ tiếng Việt bằng

JvnTextPro, thực hiện tiến hành so khớp và đối chiếu với bộ từ điển cảm xúc SO-CAL.

Trong trường hợp, những từ loại không có trong bộ từ điển, hệ thống sẽ đánh trọng số là 0. Trong trường hợp này sẽ loại bỏ các phản hồi có tổng giá trị cảm xúc là (0). Sau khi hệ thống thực hiện đánh trọng số cho các phản hồi, sẽ chọn ra những phản hồi có tổng giá trị cảm xúc trong khoảng từ [-1…0) và (0…1]. Loại bỏ những câu có giá trị bằng (0).

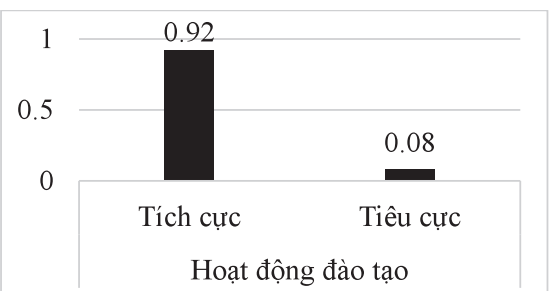
3. Thực nghiệm và đánh giá kết quả

Để thực hiện đánh giá mô hình SVM trên tập dữ liệu huấn luyện, chúng tôi sử dụng các chỉ số độ phủ (recall), độ chính xác (precision) và chỉ số cân bằng giữa 2 độ đo trên - F1 (F-measure) [5]. Để đánh giá mức độ chính xác của mô hình được huấn luyện, tiến hành chạy thực nghiệm trên tập dữ liệu như sau: Dữ liệu đầu vào của quá trình huấn luyện được cho trong bảng 4.

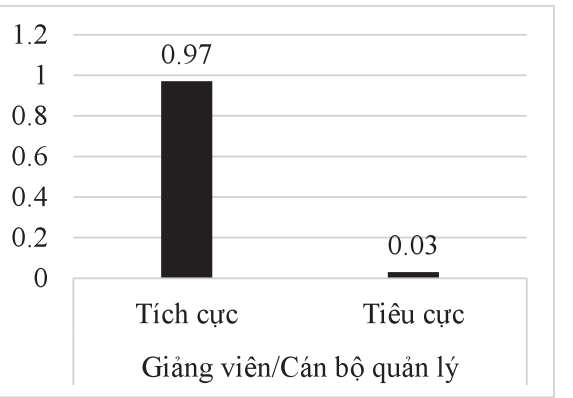
Bảng 4. Dữ liệu đầu vào quá trình huấn luyện

STT	Phân lớp	Số lượng
1	Tích cực	500
2	Tiêu cực	500

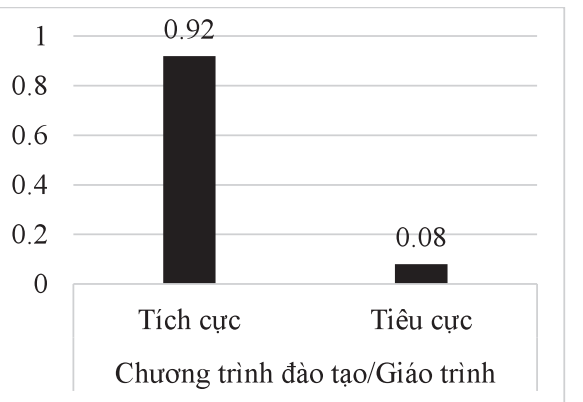
Sau khi hoàn tất quá trình huấn luyện. Tiến hành thu thập một số lượng lớn các phản hồi của đối tượng phi công quân sự tại trường Sĩ quan Không quân. Cụ thể với số lượng tập test là 800 ý kiến phản hồi về nội dung Hoạt động đào tạo, Giảng viên/Cán bộ quản lý, Chương trình đào tạo/Giáo trình, Cơ sở vật chất, thiết bị đào tạo và thư viện được thu thập trên trang thông tin phản hồi. Quá trình kiểm nghiệm được tiến hành như sau: lần lượt chọn các nội dung đưa vào phân lớp – sau đó tiến hành tính toán các độ đo. Kết quả thực nghiệm thực tế được trình bày dưới đây.



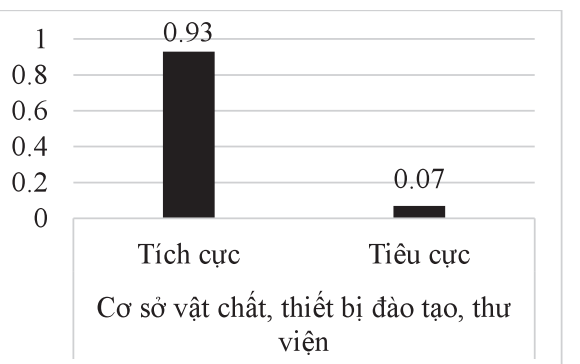
Hình 5. Tỷ lệ % phản hồi Hoạt động đào tạo



Hình 6. Tỷ lệ % phản hồi Giảng viên/Cán bộ quản lý



Hình 7. Tỷ lệ % phản hồi Chương trình đào tạo/Giáo trình



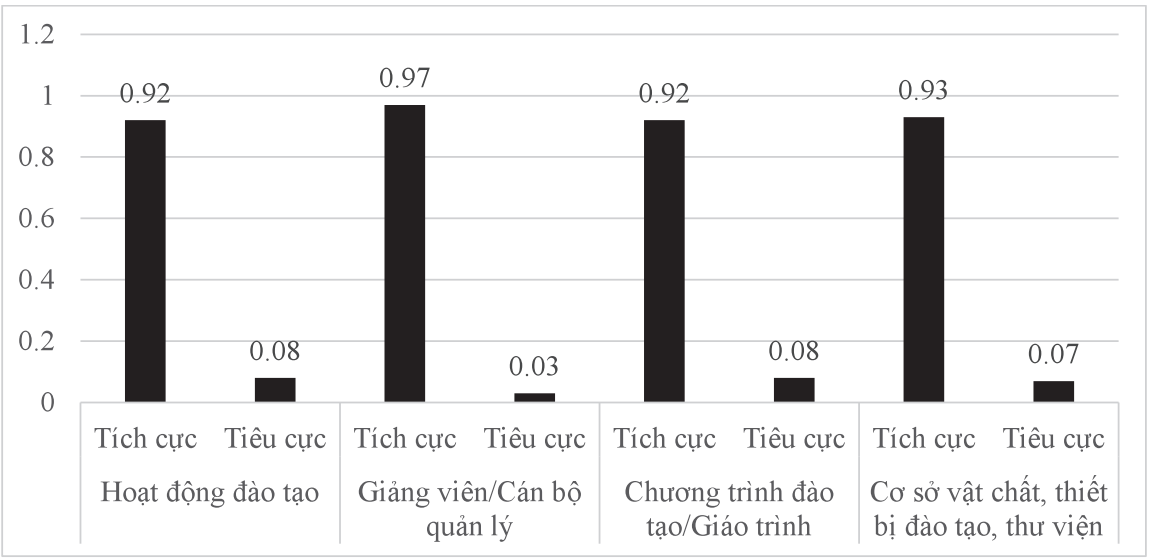
Hình 8. Tỷ lệ % phản hồi Cơ sở vật chất, thiết bị đào tạo và thư viện

Qua phân tích thực tế và kết quả chương trình, có được bảng thống kê kết quả tỷ lệ phản hồi tích cực cho từng nội dung phản hồi như bảng 5.

Bảng 5. Kết quả thực nghiệm SVM

Nội dung lấy phản hồi	Độ chính xác (Precision)	Độ phủ (Recall)	Độ đo F1 (F-measure)
Hoạt động đào tạo	0.91	0.92	0.915
Giảng viên/Cán bộ quản lý	0.94	0.97	0.955
Chương trình đào tạo/Giáo trình	0.94	0.92	0.93
Cơ sở vật chất, thiết bị đào tạo và thư viện	0.92	0.93	0.925

Tỷ lệ (%) phản hồi các nội dung được thể hiện trong hình 9 dưới đây.



Hình 9. Kết quả phản hồi tích cực các nội dung

Theo số liệu thực nghiệm trên trong các nội dung Hoạt động đào tạo, Giảng viên/Cán bộ quản lý, Chương trình đào tạo/Giáo trình, Cơ sở vật chất, thiết bị đào tạo và thư viện. Nội dung Giảng viên/Cán bộ quản lý đạt mức tích cực rất tốt. Qua số liệu phân tích cho thấy người học phản ánh thực chất và phù hợp với chất lượng giảng viên của Nhà trường ngày được nâng cao từ chuyên môn đến phương pháp giảng dạy, ứng dụng công nghệ thông tin, cập nhật kiến thức mới, sử dụng hiệu quả trang thiết bị dạy học. Đáp ứng nhu cầu ngày càng cao của người học.

4. Kết luận

Kết quả thực nghiệm cho thấy mô hình phân lớp này đã đạt được kết quả khả quan, đặc biệt là nội dung về Giảng viên/Cán bộ quản lý. Để đánh giá, phân tích chính xác hơn nữa, dự kiến sẽ thu thập và xây dựng bộ dữ liệu thử nghiệm lớn. Đồng thời, sẽ thử nghiệm áp dụng các loại mô hình khác nhau vào bài toán phân lớp để xác định loại mô hình phù hợp nhất.

Qua nghiên cứu này, chúng tôi xây dựng bộ từ điển cảm xúc tiếng Việt được thông dịch từ bộ từ điển cảm xúc tiếng Anh. Trong đó, có điều chỉnh, đánh giá qua những thử nghiệm trong thực tế. Xây dựng công cụ hỗ trợ cho việc tìm kiếm dữ liệu trên trang thông tin. Xây dựng mô hình phân tích, đánh giá phản hồi có thể ứng dụng trong thực tế.

Trong thời gian tới để tăng hiệu quả trong việc phân tích, có một số cải tiến và hướng phát triển như sau: Cập nhật, bổ sung bộ từ điển cảm xúc tiếng Việt. Xây dựng mô hình phân tích với một số giải thuật xử lý ngôn ngữ tự nhiên tiếng Việt. Làm giàu thêm bộ dữ liệu huấn luyện SVM để giúp cho việc phân lớp nâng cao độ chính xác trong phân tích phản hồi.

Tài liệu tham khảo

1. Duy, N. N., & Tươi, P. T (2016), *Tóm tắt văn bản trên cơ sở phân loại ý kiến độc giả của báo mạng tiếng Việt*, *SCIENCE & TECHNOLOGY, 19* (K5).
2. Cao, Đ. T., & Nguyên, K. P (2012), *Phân loại văn bản với máy học Vector hỗ trợ và cây quyết định*, Tạp chí Khoa học Trường Đại học Cần Thơ, (21a), 52-63.
3. Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020), *An introduction to machine learning*, Clinical Pharmacology & Therapeutics, *107*(4), 871-885.
4. Phu, V. N., & Tuoi, P. T (2014, October), *Sentiment classification using enhanced contextual valence shifters*, In 2014 International Conference on Asian Language Processing (IALP) (pp. 224-229), IEEE.
5. Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-49).
6. Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.
7. Nguyễn Việt Khoa (2012), VIETTIEN Dictionary for Mac, <https://nguyenvietkhoa.edu.vn/viettien-dic/, truy cập ngày 28/4/2022>

ANALYSIS STUDENTS' FEEDBACK BASED ON EMOTIONAL CLASSIFICATION METHODS

Hoang Ngoc Duong

Air Force Officers College

Abstract: *In this study, we build a model to analyze the learners' feedback through automatically classifying and labeling the feedback. The main work includes the following steps: Build a tool to get data from the feedback page, clean the data, build a data classification model based on the learner feedback set. Then conduct an analysis based on the Vietnamese emotion dictionary. The classification and prediction of labels is done based on the Support Vector Machine (SVM) method. The experiment gave positive results on the feedback set on some contents of Training activities, Lecturers/ Managers, Training Programs/ Textbooks, Facilities, training equipment and library.*

Keywords: *Text classification; Learner's opinion; Emotion dictionary; Support Vector Machine*