HUTECH
Đại học Công nghệ Tp.HCM

HỘI ĐỒNG BẢO VỆ LUẬN VĂN THẠC SĨ

Chuyên ngành: **CÔNG NGHỆ THÔNG TIN**

Mã ngành: **8 48 02 01**

HUTECH
Đại học Công nghệ Tp.HCM

CHUYÊN NGÀNH CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI:

ỨNG DỤNG KHAI THÁC DỮ LIỆU TRONG LĨNH VỰC GIÁO DỤC

CÁN BỘ HƯỚNG DẪN: TS LÊ THỊ NGỌC THƠ

HỌC VIÊN THỰC HIỆN: VÕ MINH QUÂN

TP.HCM, 17/01/2021

Nội Dung Trình Bày

1. GIỚI THIỆU ĐỀ TÀI
2. NỘI DUNG NGHIÊN CỨU
3. THỰC NGHIỆM VÀ ĐÁNH GIÁ
4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. GIỚI THIỆU ĐỀ TÀI

- Ngày nay việc thu thập ý kiến, cảm xúc phản hồi đánh giá của con người trong trong các vấn đề là một việc rất phổ biến mà dựa vào đó ta có thể đưa ra những đánh giá, nhận xét cho các vấn đề liên quan.
- Một vài lĩnh vực phổ biến cho việc thu thập và sử dụng ý kiến phản hồi như :
  - Kinh nghiệm cá nhân và ý kiến đánh giá, diễn đàn, blog, v.v.
  - Nhận xét về bài viết, vấn đề, chủ đề, bài đánh giá, v.v.
  - Thông tin phản hồi sản phẩm tại các trang bán hàng trực tuyến

1. GIỚI THIỆU ĐỀ TÀI

Ứng dụng của việc phân tích ý kiến

❖ Các doanh nghiệp và tổ chức

- Các sản phẩm và dịch vụ , thông tin thị trường.
- Các doanh nghiệp tìm kiếm ý kiến của người tiêu dùng bằng cách sử dụng tư vấn, khảo sát và nhóm tập trung, v.v.

❖ Cá nhân

- Ra quyết định mua sản phẩm hoặc sử dụng dịch vụ.
- Tìm ý kiến công chúng về các ứng cử viên và vấn đề chính trị.

1. GIỚI THIỆU ĐỀ TÀI

Lý do chọn đề tài

❖ Việc khai thác dữ liệu trong lĩnh vực giáo dục gần đây chú trọng và quan tâm góp phần lớn cải thiện chất lượng giáo dục.

❖ Từ những thực tế ở trường Đại học Công Nghệ TP.HCM là việc phân tích đánh giá ý kiến khảo sát sinh viên về chất lượng giảng dạy mỗi học kỳ đều được làm thủ công. Vì thế nhu cầu về một hệ thống phân tích ý kiến đánh giá tự động và hiệu quả là có thật.

1. GIỚI THIỆU ĐỀ TÀI

Mục tiêu luận văn

- Tìm hiểu về các phương pháp phân tích ý kiến (đưa về bài toàn phân lớp dữ liệu).

- Giải quyết bài toán phân tích ý kiến khảo sát chất lượng giảng dạy của giảng viên tại Trường Đại học Công Nghệ TP.HCM.

- So sánh độ hiệu quả của các phương pháp phân lớp khác nhau trên bài toán phân tích ý kiến khảo sát chất lượng giảng dạy.

2. NỘI DUNG NGHIÊN CỨU

Phân tích ý kiến

Phân tích ý kiến được chia làm 4 hướng nghiên cứu chính:

- ❖ Phân lớp chủ quan.
- ❖ **Phân lớp cảm xúc.**
- ❖ Tóm tắt ý kiến.
- ❖ Khai thác ý kiến trên đặc trưng.

2. NỘI DUNG NGHIÊN CỨU

Phân tích cảm xúc

❖ Phân tích cảm xúc (Sentiment analysis) là nhằm phát hiện ra thái độ, màu sắc tình cảm, khuynh hướng niềm tin trong một vấn đề nào đó.

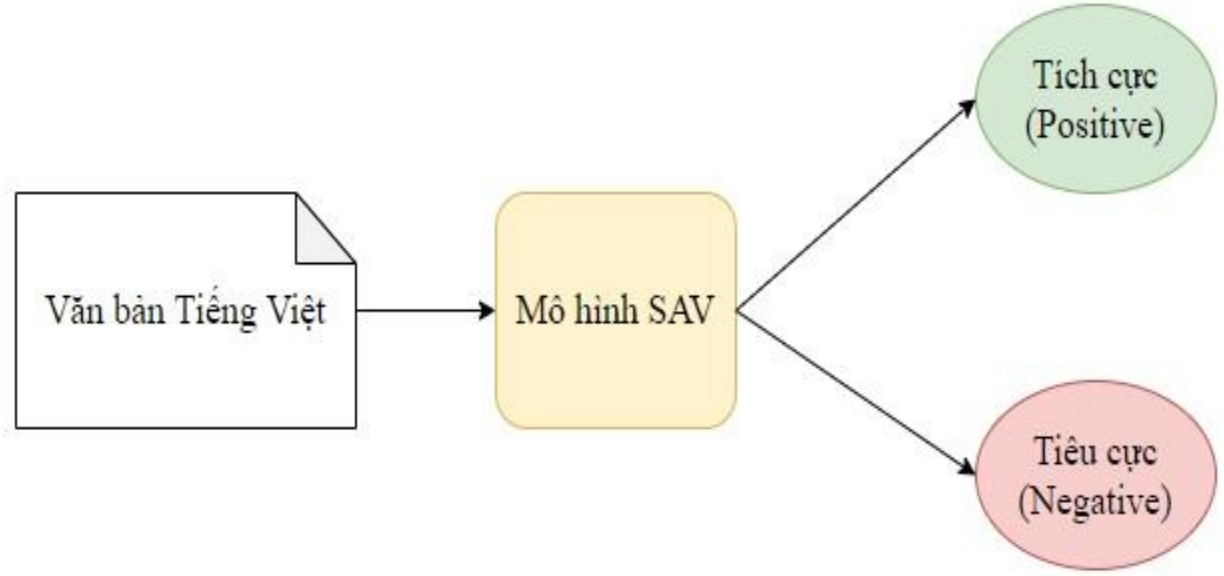
❖ Bài toán phân tích cảm xúc là bài toán dạng phân lớp cảm xúc dựa trên văn bản ngôn ngữ tự nhiên.

Phân tích cảm xúc (tiếp theo)

Bài toán phân tích cảm xúc thường được phân thành các bài toán có độ khó như sau:

- ❖ Đơn giản: Phân tích cảm xúc thành 2 lớp là tích cực (positive) và tiêu cực (negative).
- ❖ Trung bình: Xếp hạng cảm xúc theo mức độ.
- ❖ Khó: Phát hiện mục tiêu nguồn gốc của cảm xúc hoặc các loại cảm xúc phức tạp.

Phân tích cảm xúc (tiếp theo)



Mô hình xử lý Sentiment Analysis Vietnamese (SAV).

Phân tích cảm xúc (tiếp theo)

Hiện nay bài toán phân tích cảm xúc có thể được giải quyết dựa trên những phương pháp như:

- ❖ Theo phương pháp phân lớp không giám sát.
- ❖ **Theo phương pháp phân lớp có giám sát.**
- ❖ Phân tích cảm xúc dựa trên khía cạnh.
- ❖ Phân loại cảm xác dựa trên chủ đề.

2. NỘI DUNG NGHIÊN CỨU

Một số phương pháp phân lớp

- **Phương pháp phân lớp Naïve Bayes**
- **Phương pháp phân lớp SVM (Support Vector Machines)**
- Phương pháp K-Nearest Neighbor
- Phương pháp Phương pháp Linear Least Square Fit (LLSF)
- Phương pháp Entropy cực đại

Phương pháp biểu diễn văn bản

❖ Biểu diễn văn bản là một bước quan trọng trong khai thác dữ liệu văn bản, truy vấn thông tin và xử lý ngôn ngữ tự nhiên.

❖ Các mô hình biểu diễn văn bản truyền thống như mô hình túi từ (bag-of-word), mô hình không gian vector là các mô hình thường được sử dụng nhất

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Dữ liệu thực nghiệm

- Dữ liệu thực nghiệm được tổng hợp từ ý kiến đánh giá chất lượng giảng viên trong học kì I năm học 2016-2017 của Trường Đại học Công Nghệ TP. Hồ Chí Minh.

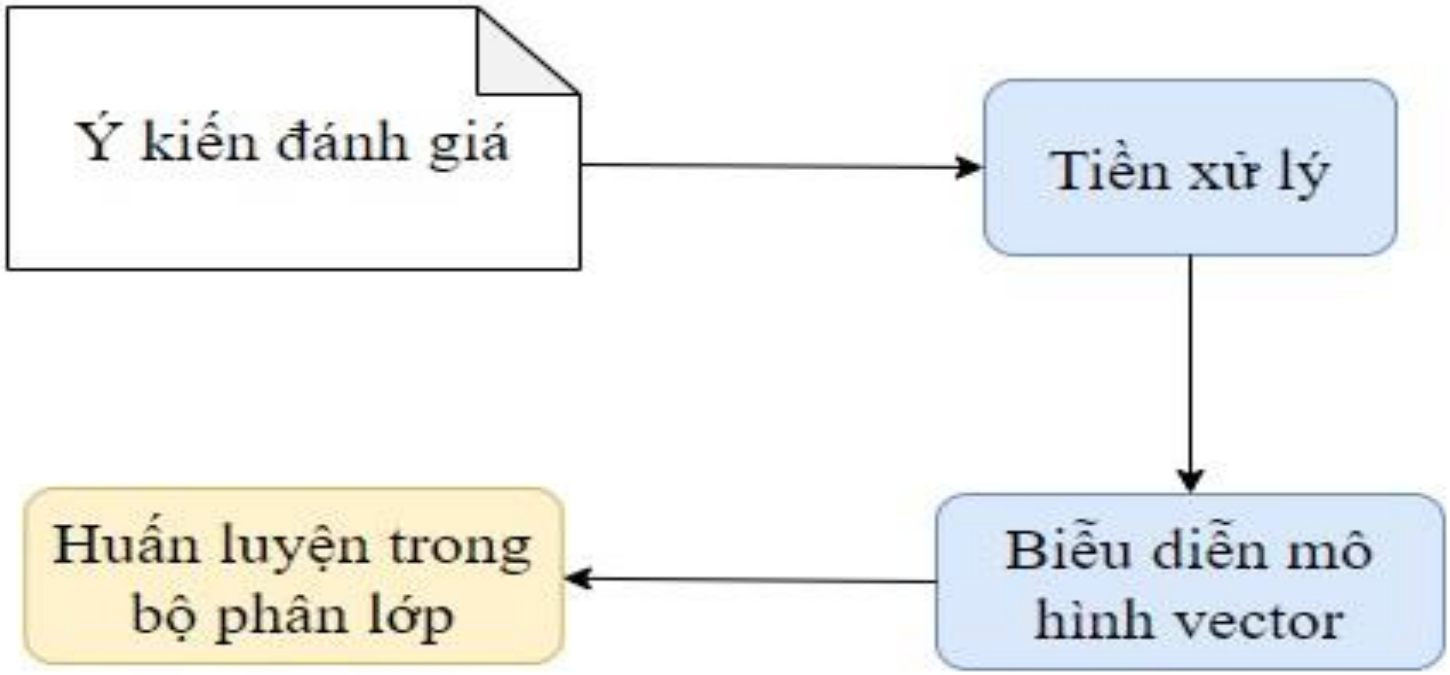
- Tập dữ liệu trích xuất gồm 1.000 dữ liệu trong đó bao gồm 500 dữ liệu thể hiện ý kiến tích cực (positive) và 500 dữ liệu tiêu cực (negative).

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Bộ phân lớp cảm xúc sẽ chia nhỏ thành hai giai đoạn bao gồm:

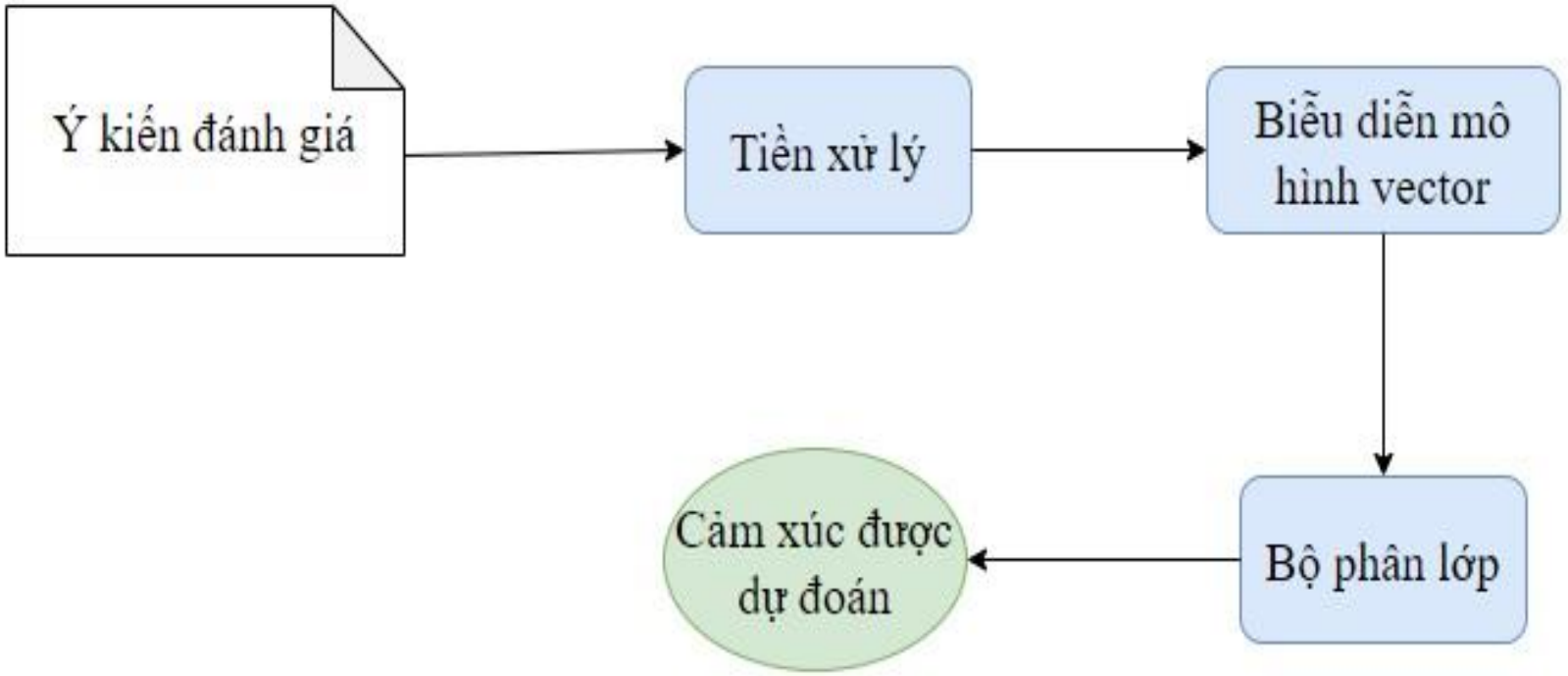
- ❖ Giai đoạn huấn luyện mô hình (training),
- ❖ Giai đoạn kiểm tra mô hình (test)

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Quy trình huấn luyện bộ phân lớp.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Quy trình kiểm tra bộ phân lớp.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

giảng_viên dạy dễ hiểu trừ điểm khá gắt nghỉ một buổi trừ điểm trong khi phải học buổi	tieu_cuc
thầy gắt quá cho tập_thể_lực xong là không học nổi nữa	tieu_cuc
dạy toàn lên đứng nói một_mình không quan_tâm sinh_viên bắt sinh_viên làm theo như_khỉ	tieu_cuc
thấy rất nhiệt_tình và vui_tính	tich_cuc
cô có_thể điểm_danh thư_thả thời_gian cho sinh_viên cũng vì nhiều lí_do khác nhau mà nhiều sinh_viên không_thể đến đúng	tieu_cuc
sắp_xếp lịch bù khá nhiều nhờ giảng_viên khác dạy thế	tieu_cuc

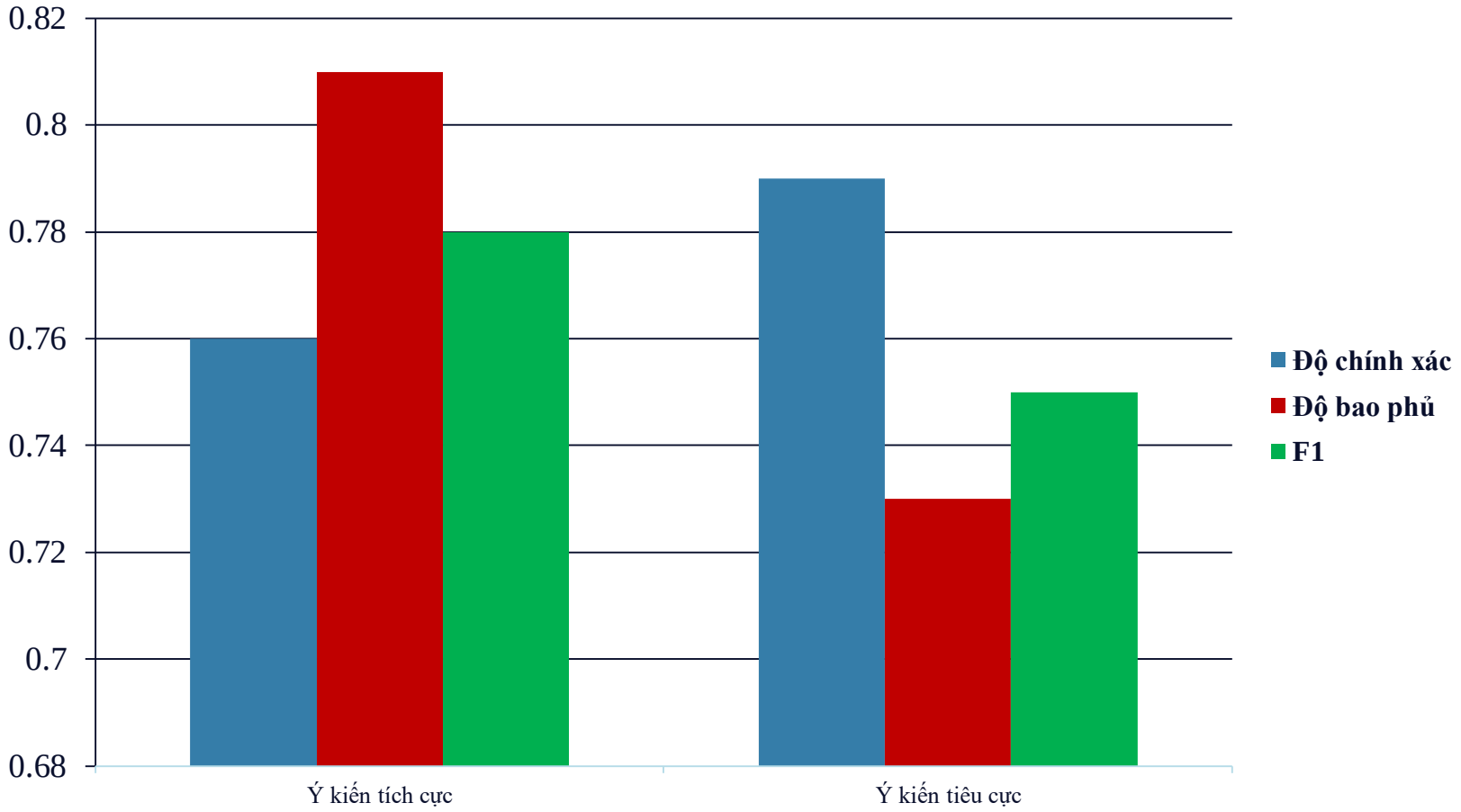
Dữ liệu sau khi được tiền xử lý và gán nhãn.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Lần chạy	Thuật toán	Phân lớp tích cực			Phân lớp tiêu cực			Độ chính xác	Độ bao phủ	F1
		Độ chính xác	Độ bao phủ	F1	Độ chính xác	Độ bao phủ	F1			
1	SVM	0.66	0.89	0.75	0.87	0.62	0.72	0.76	0.75	0.74
2	SVM	0.77	0.76	0.76	0.77	0.78	0.78	0.77	0.77	0.77
3	SVM	0.82	0.77	0.8	0.76	0.81	0.78	0.79	0.79	0.79
4	SVM	0.71	0.83	0.76	0.82	0.7	0.76	0.77	0.76	0.76
5	SVM	0.8	0.77	0.78	0.72	0.75	0.73	0.76	0.76	0.76
6	SVM	0.78	0.82	0.8	0.8	0.76	0.78	0.79	0.79	0.79
7	SVM	0.75	0.87	0.8	0.87	0.74	0.8	0.81	0.8	0.8
8	SVM	0.73	0.73	0.73	0.71	0.71	0.71	0.72	0.72	0.72
9	SVM	0.77	0.86	0.81	0.77	0.64	0.70	0.77	0.75	0.77
10	SVM	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
Trung bình		0.76	0.81	0.78	0.79	0.73	0.75	0.77	0.77	0.77

Kết quả thực nghiệm bộ phân lớp với SVM

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Kết quả thực nghiệm phân lớp cảm xúc

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Lần chạy	Thuật toán	P	R	F1
1	SVM	0.77	0.77	0.77
2	SVM	0.77	0.77	0.77
3	SVM	0.77	0.77	0.77
4	SVM	0.77	0.77	0.76
5	SVM	0.77	0.77	0.77

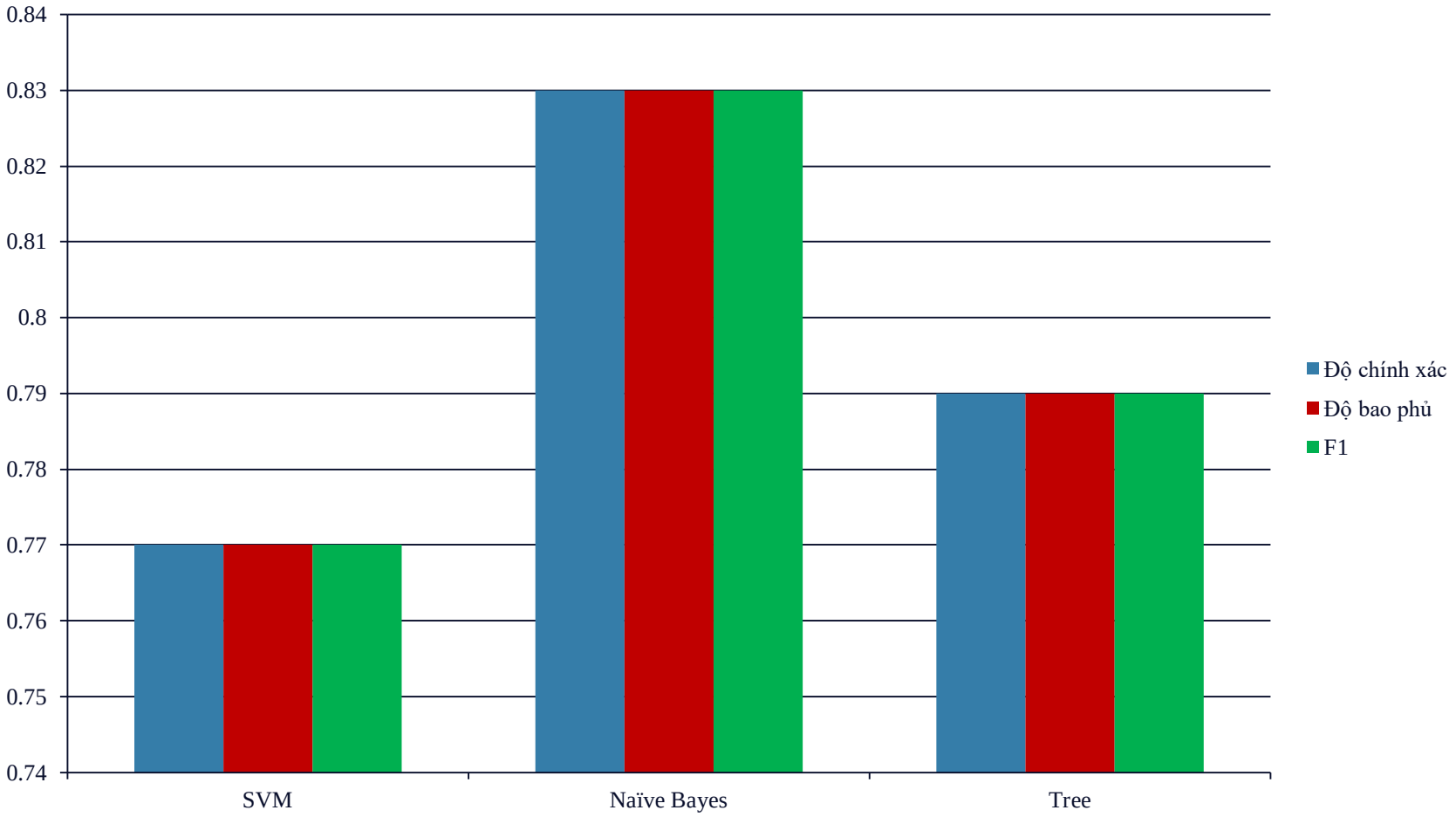
Kết quả thực nghiệm trong 5 lần chạy

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Phương pháp	Độ chính xác	Độ bao phủ	F1
SVM	0.77	0.77	0.77
Naïve Bayes	**0.83**	**0.83**	**0.83**
Decision Tree	0.79	0.79	0.79

So sánh độ hiệu quả giữa các phương pháp phân lớp

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ



Bảng so sánh độ hiệu quả giữa các phương pháp phân lớp

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Kết quả đạt được

- Xây dựng được một bộ phân lớp ý kiến đánh giá với độ chính xác lên tới **83%.**

- So sánh một số phương pháp phân lớp trên cùng tập dữ liệu từ đó làm cơ sở lý thuyết tham khảo cho các nghiên cứu liên quan.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết quả đạt được

- Xây dựng thành công mô hình dự đoán ý kiến đánh giá trong lĩnh vực giáo dục. Độ chính xác của mô hình lên đến **83%** với phương pháp phân lớp Naïve Bayes**.**

- So sánh độ hiệu quả giữa các phương pháp phân lớp với nhau trên cùng tập dữ liệu làm nguồn tài liệu tham khảo cho các nghiên cứu liên quan.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hạn chế

- Chưa phân loại được các ý kiến mang ý kiến trung tính.
- Mô hình vẫn phụ thuộc vào việc lọc và gán nhãn dữ liệu thủ công.
- Việc biểu diễn văn bản thành vector chưa xét đến ngữ nghĩa trong câu.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hướng phát triển

- Tăng số lượng dữ liệu huấn luyện để cải thiện độ chính xác phân lớp.
- Cải tiến phương pháp biểu diễn văn bản thành vector, cũng như giảm chiều vector.
- Thử nghiệm các phương pháp phân lớp mới.

4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hướng phát triển

Ngoài ra luận văn có thể mở rộng và phát triển ở các hướng sau:

- ❖ Tăng số lớp dự đoán cảm xúc lên, tự động nhận diện các ý kiến không mang cảm xúc.
- ❖ Kết hợp nhiều phương pháp phân lớp khác nhau để nâng cao độ chính xác.

Cảm ơn quý thầy cô và các bạn đã theo dõi lắng nghe