HUTECH
Đại học Công nghệ Tp.HCM
475A Dien Bien Phu Street, Ward 25, Binh Thanh Dítrict
Tel: +84 8 5445 7777, Fax: +84 8 5445 4444
UNIVERSITY OF TECHNOLOGY

Đề tài:

PHÂN LOẠI VĂN BẢN DỰA TRÊN MÔ HÌNH ĐỒ THỊ

Học viên thực hiện: HOÀNG NGỌC DƯƠNG
Cán bộ hướng dẫn: PGS.TS VÕ ĐÌNH BẢY

NỘI DUNG BÁO CÁO

1 GIỚI THIỆU TỔNG QUAN

2 BÀI TOÁN PHÂN LOẠI VĂN BẢN

3 THỰC NGHIỆM, ĐÁNH GIÁ

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

GIỚI THIỆU TỔNG QUAN

- Phân loại văn bản là quá trình gán các văn bản vào một hay nhiều lớp văn bản đã được xác định từ trước.
- Có thể phân loại các văn bản một cách thủ công, tức là đọc nội dung từng văn bản và gán nó vào một loại nào đó.
- Hệ thống quản lý tập gồm nhiều văn bản cho nên cách này sẽ tốn nhiều thời gian, công sức và do đó là không khả thi.

GIỚI THIỆU TỔNG QUAN

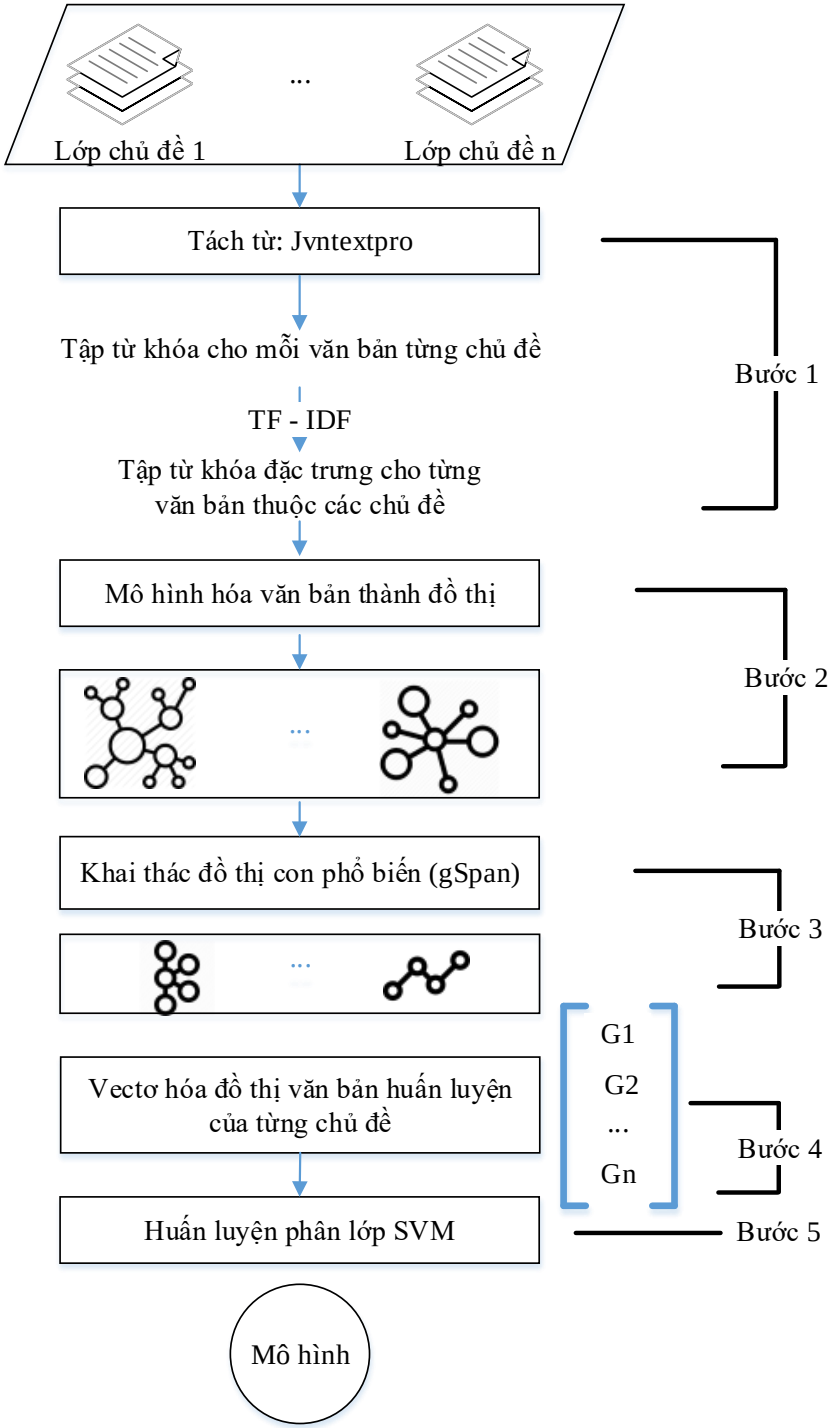
- Để phân loại có nhiều phương pháp học máy trong trí tuệ nhân tạo như: máy vector hỗ trợ (SVM), Cây quyết định, Naïve Bayes, K láng giềng gần nhất ...
- Đối với những phương trên thường chuyển mô hình văn bản thành dạng véc tơ. Nên độ lớn, số chiều của véc tơ lớn nên việc xử lý chậm.
- Luận văn sử dụng mô hình đồ thị, khai thác dữ liệu đồ thị để chuyển văn bản sang dạng véc tơ.

Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Các bước thực hiện:

- Bước 1: Thực hiện việc tách từ và tính TF – IDF
- Bước 2: Mô hình hóa văn bản thành đồ thị
- Bước 3: Khai thác đồ thị con phổ biến bằng thuật toán gSpan
- Bước 4: Vec tơ hóa đồ thị văn bản
- Bước 5: Huấn luyện phân loại bằng SVM

Mô hình huấn luyện phân loại văn bản



Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

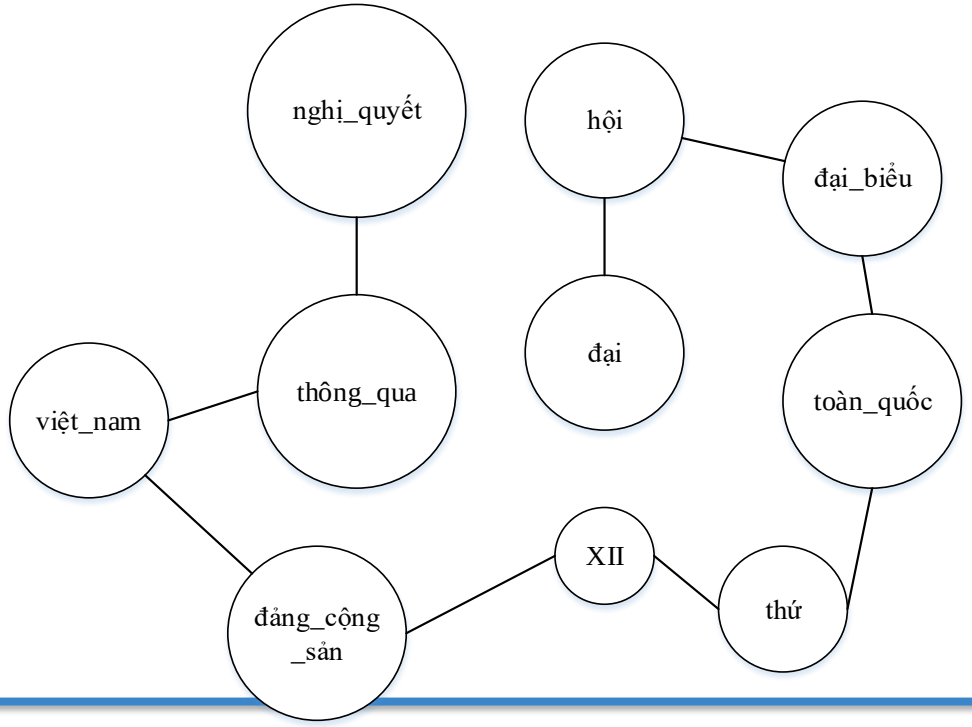
Bước 1: Thực hiện việc tách từ và tính TF – IDF

- Sử dụng bộ thư viện tách từ Jvntextpro được phát triển bởi tác giả Nguyễn Cẩm Tú.
- Tính tần suất xuất hiện của các từ trong tập văn bản bằng phương pháp TF - IDF.
- Loại bỏ những từ có tần suất xuất hiện thấp thu được bộ từ khóa cho từng văn bản.

Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Bước 2: Mô hình hóa văn bản thành đồ thị

Ví dụ ta có văn bản sau: Đại hội đại biểu toàn quốc lần thứ XII Đảng Cộng sản Việt Nam đã thông qua Nghị quyết.

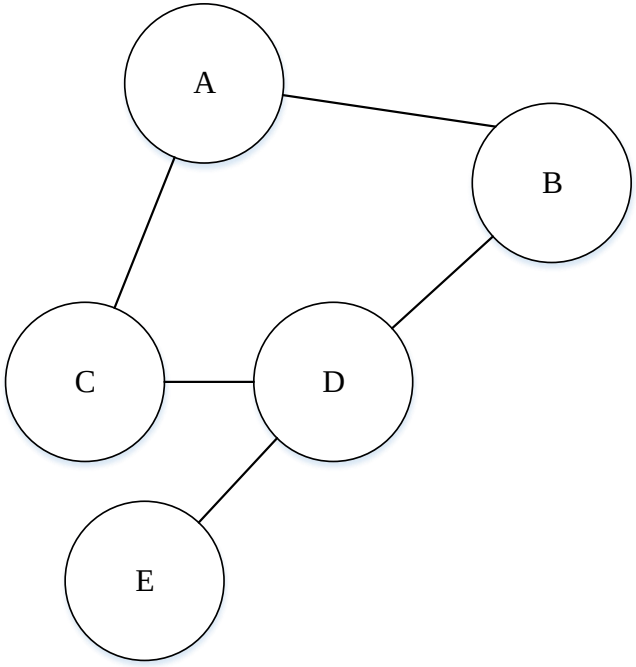


Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Bước 3: Khai thác đồ thị con phổ biến bằng thuật toán gSpan

Trong từng chủ đề, chúng ta tìm tập đồ thị con phổ biến có tần số xuất hiện lớn hơn ngưỡng phổ biến tối thiểu minsup.

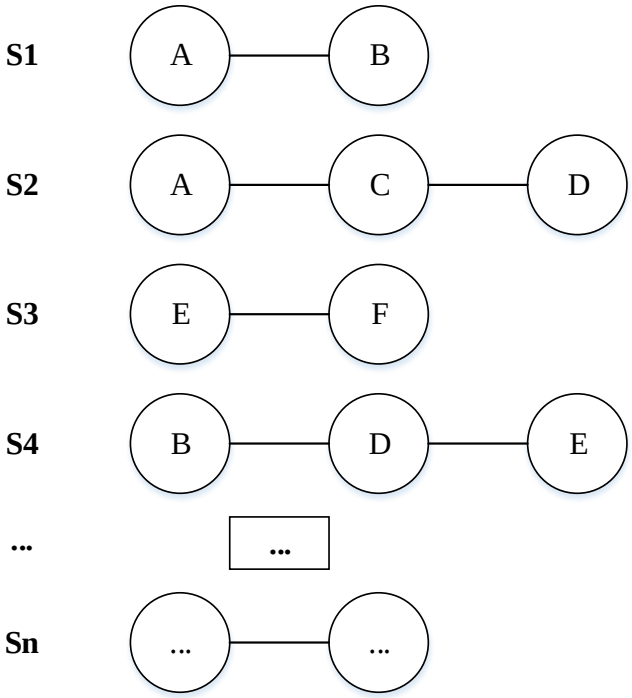
Bước 4: Vec tơ hóa đồ thị văn bản



Đồ thị văn bản G1

Vec tơ đồ thị văn bản G1 thu được
[1:1 2:1 3:0 4:1 ... n:?]

Tập đồ thị con phổ biến S = {S1, S2, ... , Sn}

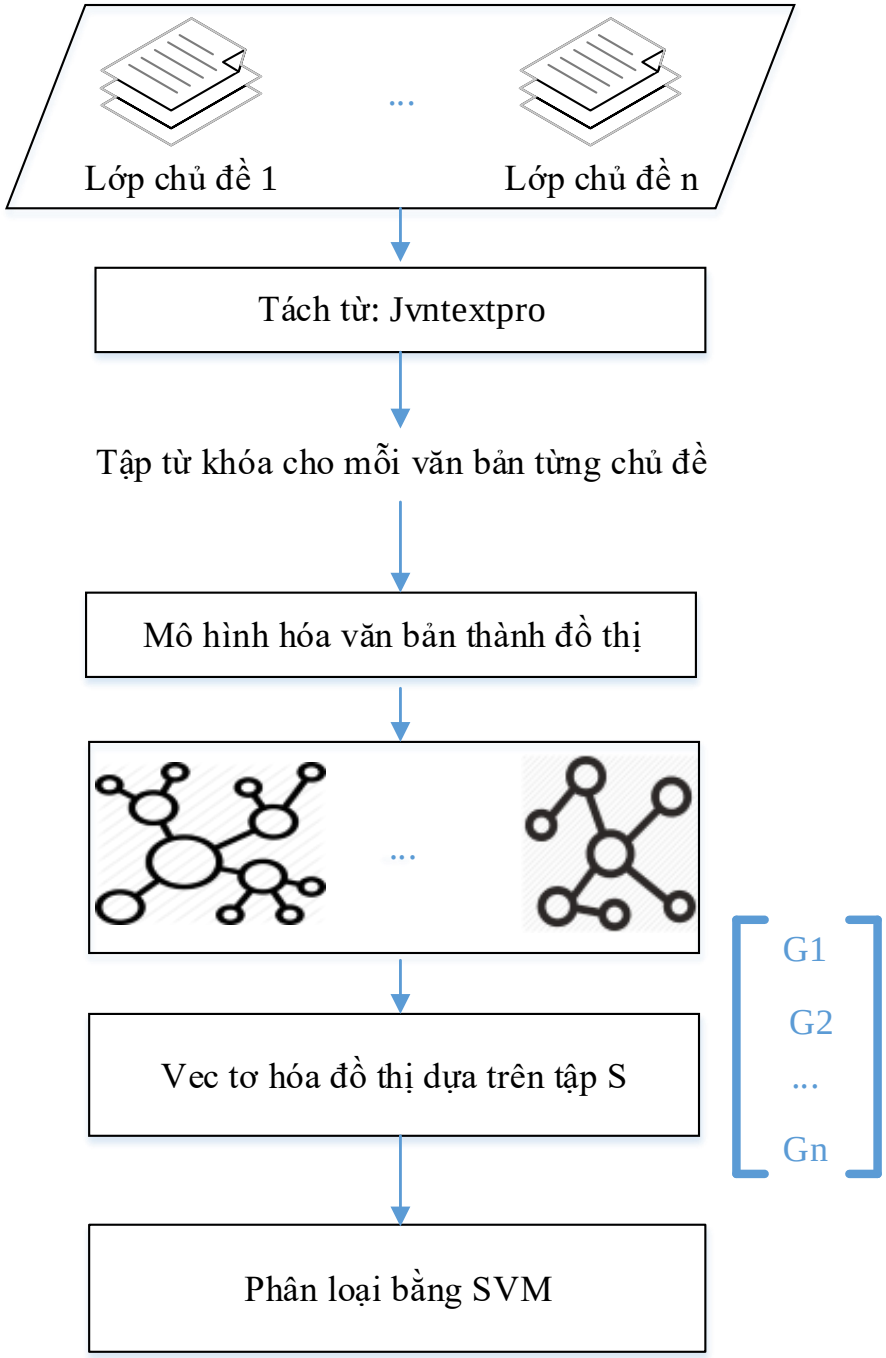


Huấn luyện phân loại văn bản dựa trên mô hình đồ thị

Bước 5: Huấn luyện phân loại bằng SVM

- SVM là một phương pháp học có giám sát để phân lớp dữ liệu.
- Công cụ khá mạnh trong nhiều lĩnh vực như: nhận dạng chữ viết, nhận dạng mặt người, phân loại văn bản, ...
- Ưu điểm của SVM là khả năng phân lớp với độ chính xác cao.

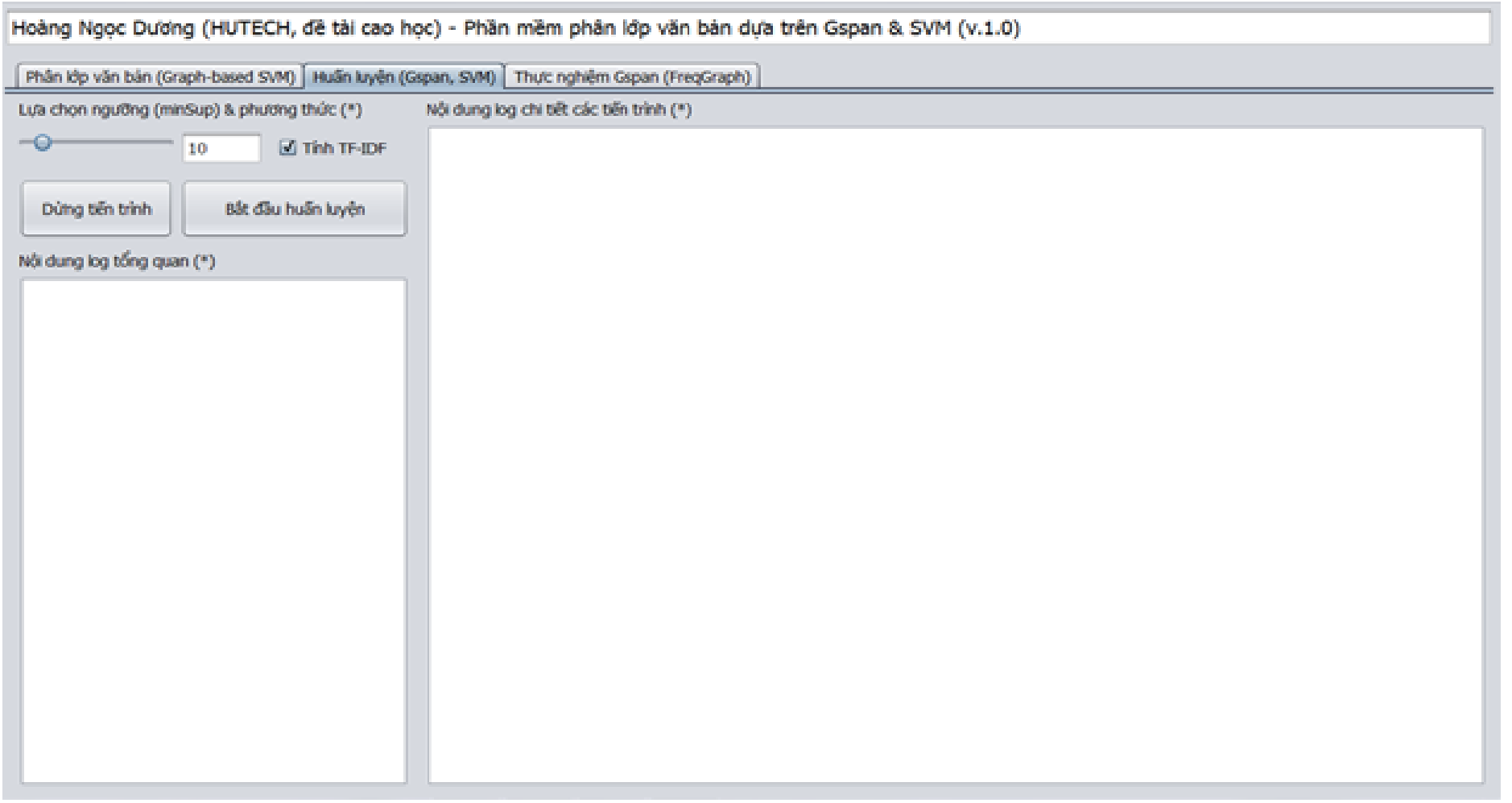
Mô hình phân loại văn bản



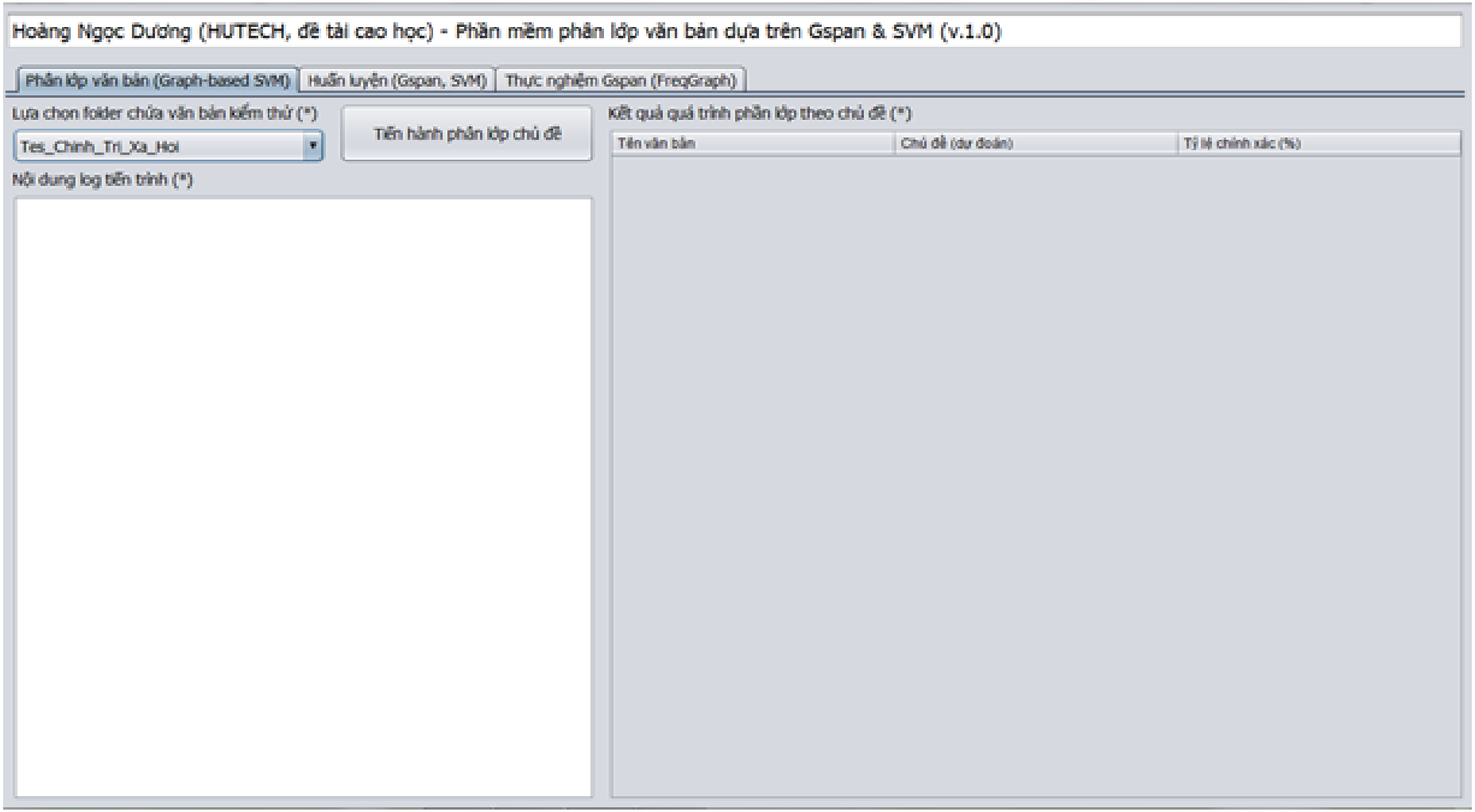
Phân loại văn bản dựa trên mô hình đồ thị

- Bước 1: Thực hiện công việc tách từ bằng bộ thư viện Jvntextpro.
- Bước 2: Mô hình hóa văn bản thành đồ thị, vec tơ hóa văn bản bằng cách chiếu lên không gian đặc trưng S (như ở pha huấn luyện) để nhận được các vec tơ đặc trưng tương ứng.
- Bước 3: Đưa qua bộ phân loại SVM đã được huấn luyện để phân loại văn bản.

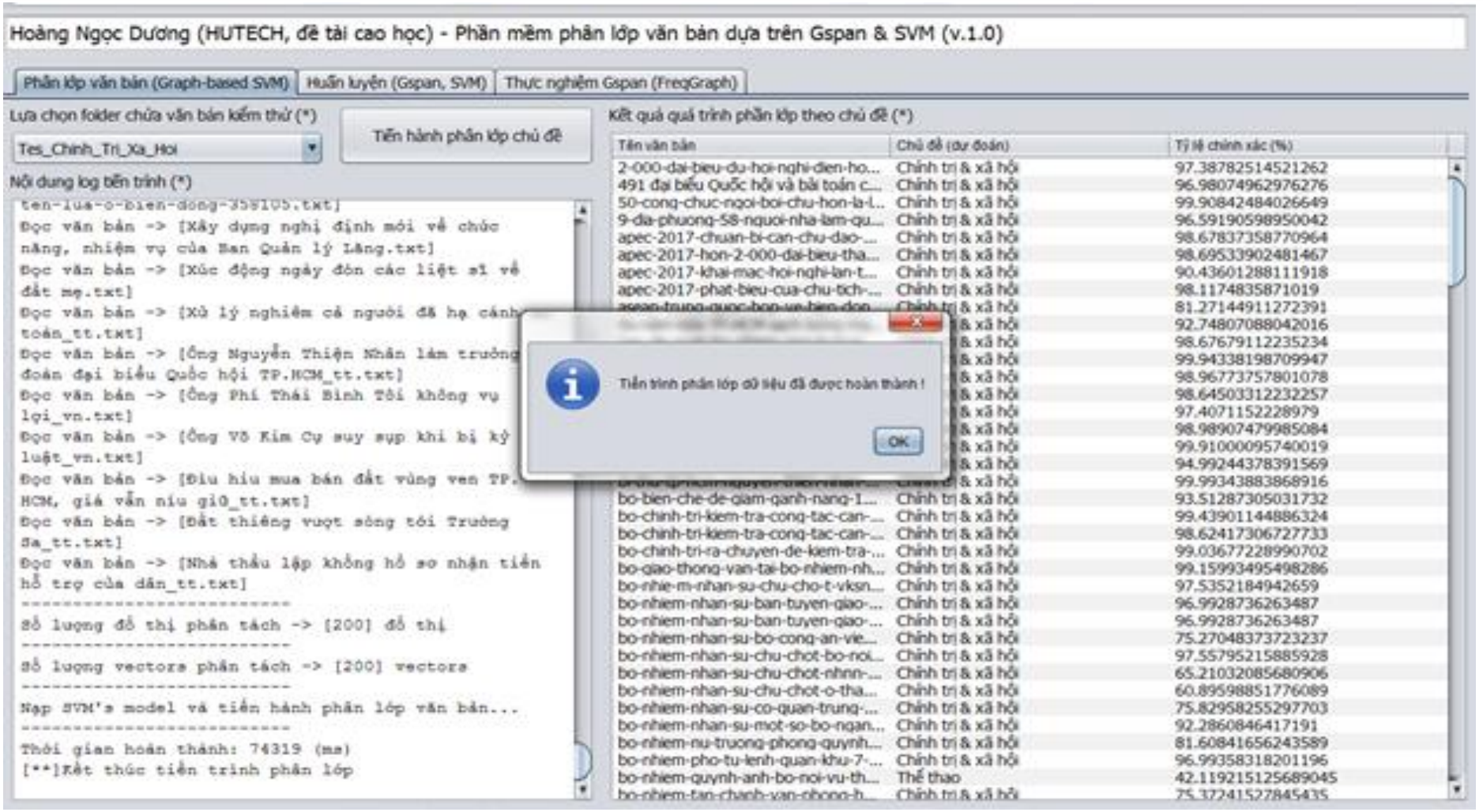
GIAO DIỆN CHƯƠNG TRÌNH



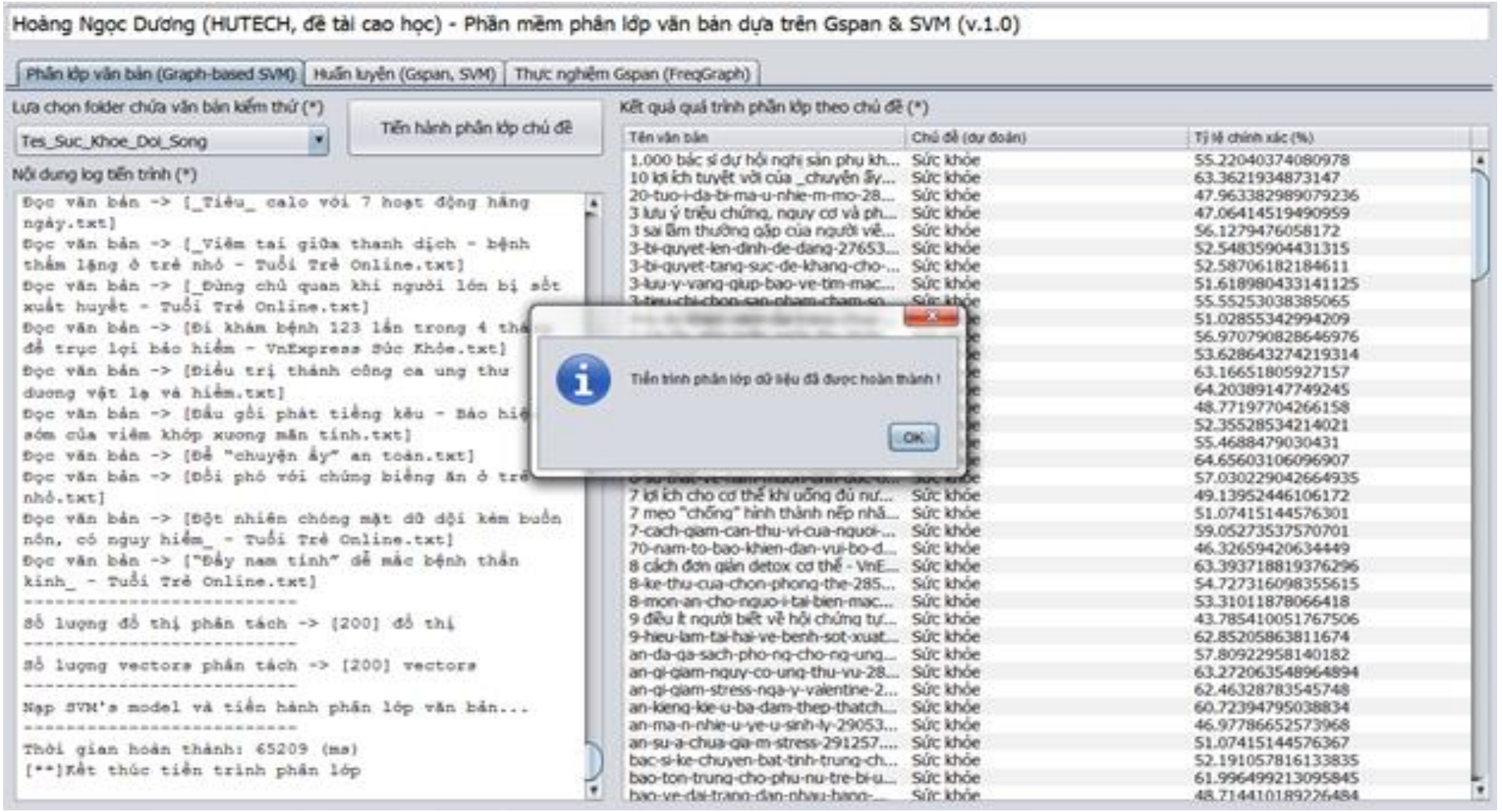
GIAO DIỆN CHƯƠNG TRÌNH



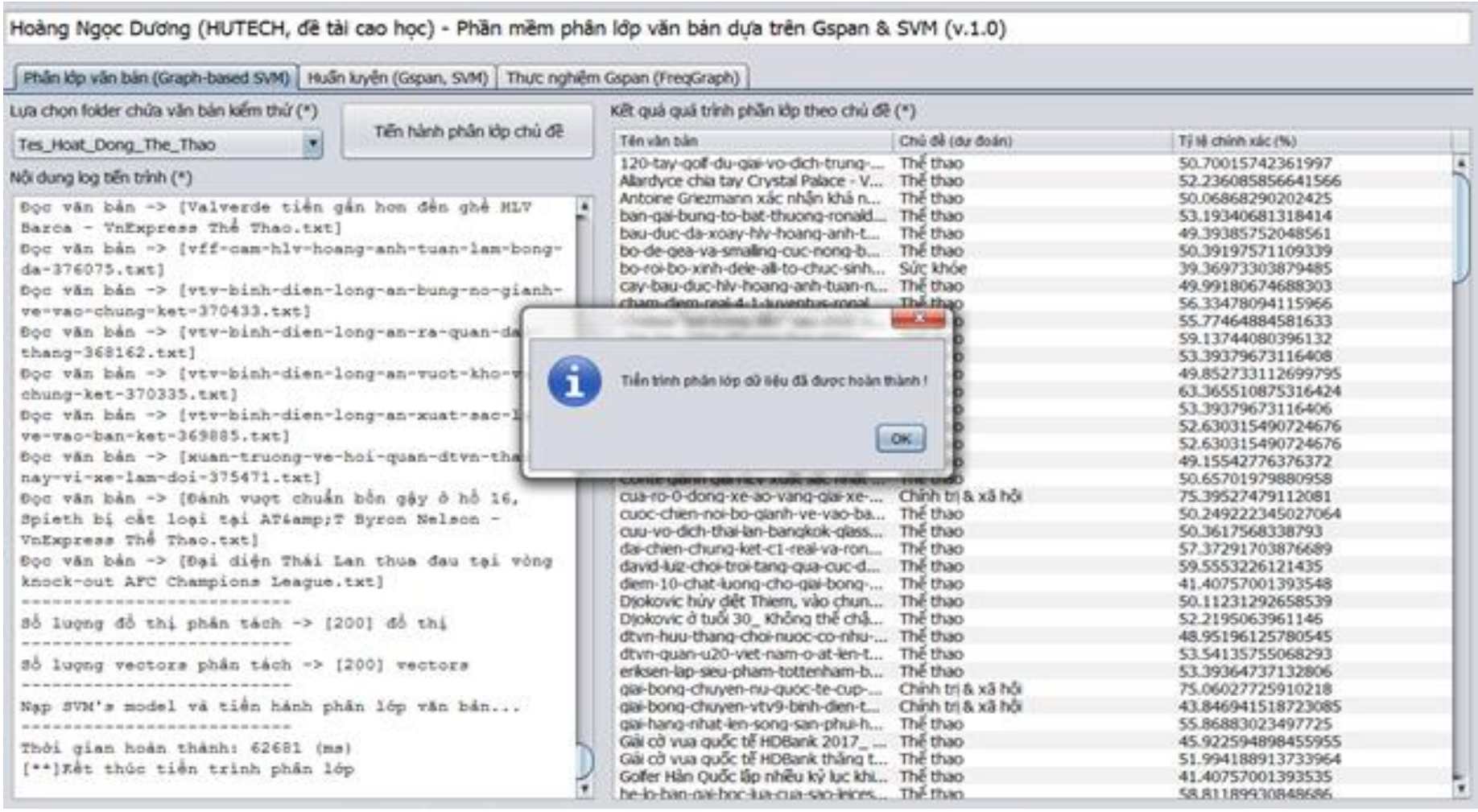
GIAO DIỆN CHƯƠNG TRÌNH



GIAO DIỆN CHƯƠNG TRÌNH



GIAO DIỆN CHƯƠNG TRÌNH



THU THẬP DỮ LIỆU

Chúng tôi tiến hành thu thập một số lượng lớn bài báo thuộc ba chủ đề: Chính trị - xã hội, Sức khỏe, Thể thao từ các nguồn tin tức điện tử: http://vnexpress.net/, http://dantri.com.vn/, http://tuoitre.vn/

THỰC NGHIỆM VÀ ĐÁNH GIÁ

1. Thực nghiệm giảm số lượng đồ thị con phổ biến thông qua TF - IDF

Chủ đề	Số văn bản	Số lượng đồ thị phổ biến (FreqGraph)								
		minSup = 20%			minSup = 30%			minSup = 40%		
		Không TF-IDF	Có TF-IDF	%	Không TF-IDF	Có TF-IDF	%	Không TF-IDF	Có TF-IDF	%
Chính trị - xã hội	200	281	68	24.20	214	47	21.96	172	37	21.51
	300	402	94	23.38	294	75	25.51	235	59	25.11
	400	613	140	22.84	476	96	20.17	380	68	17.89
Sức khỏe	200	187	58	31.02	134	46	34.33	107	36	33.64
	300	354	97	27.40	281	77	27.40	225	58	25.78
	400	567	113	19.93	432	102	23.61	346	97	28.03
Thể thao	200	234	79	33.76	192	63	32.81	157	48	30.57
	300	456	85	18.64	378	68	17.99	297	52	17.51
	400	546	122	22.34	436	98	22.48	348	72	20.69

2. Thực nghiệm mức độ chính xác của phân lớp

Dữ liệu huấn luyện:

Tên chủ đề	Số lượng văn bản
Chính trị - xã hội	300
Sức khỏe	300
Thể thao	300

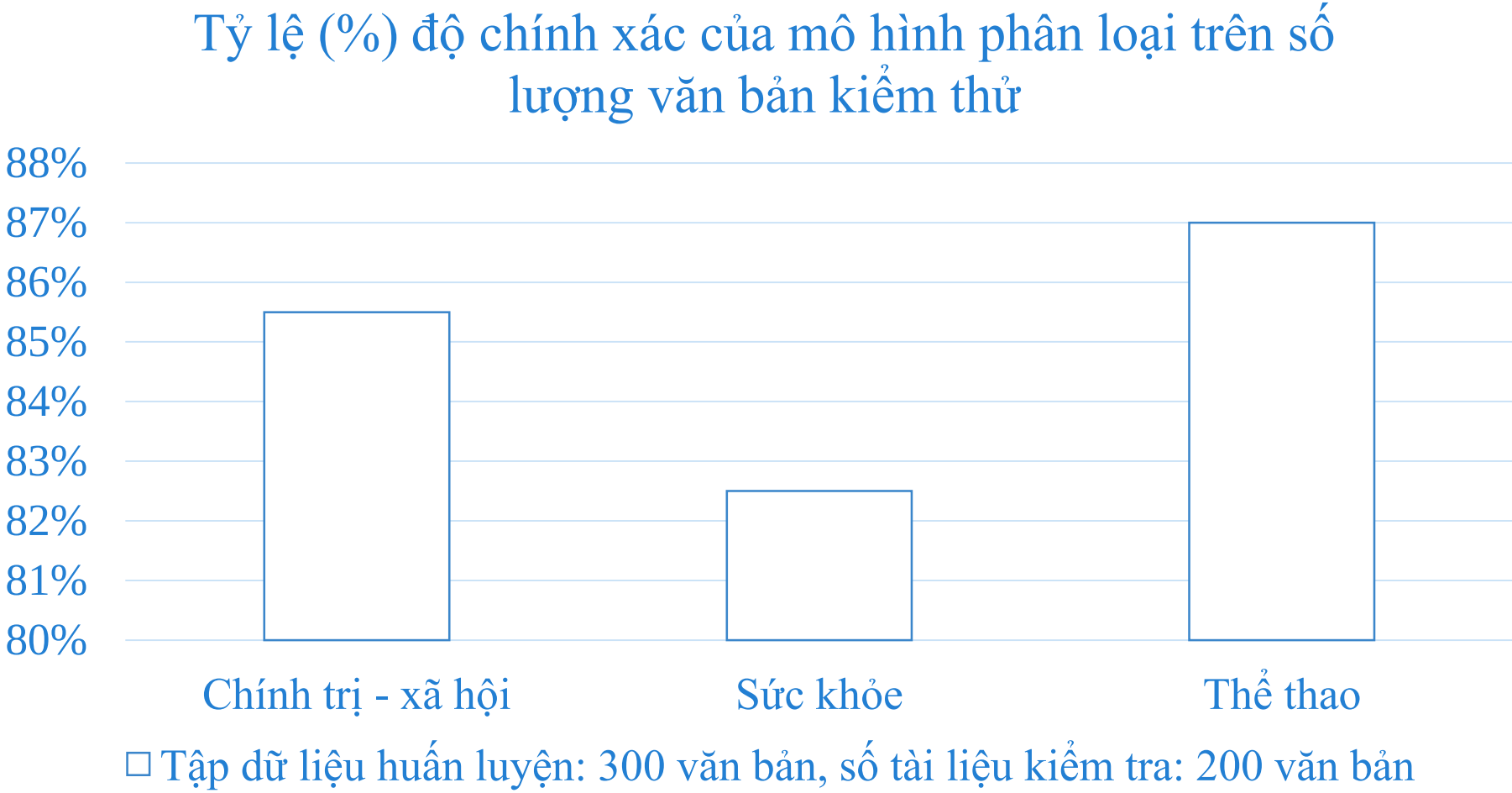
KẾT QUẢ THỰC NGHIỆM

Tên chủ đề	Độ chính xác (Precision)	Độ phủ (Recall)	Độ đo F1 (F-measure)
Chính trị - xã hội	0.834	0.855	0.844
Sức khỏe	0.917	0.825	0.868
Thể thao	0.809	0.87	0.839
Trung bình	0.853	0.850	0.850

KẾT QUẢ THỰC NGHIỆM

Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao
Chính trị - xã hội	171	8	21
Sức khỏe	15	165	20
Thể thao	19	7	174

KẾT QUẢ THỰC NGHIỆM



THỰC NGHIỆM VÀ ĐÁNH GIÁ

Thực nghiệm mức độ chính xác của phân lớp khi tăng dữ liệu huấn luyện:

Tên chủ đề	Số lượng văn bản
Chính trị - xã hội	500
Sức khỏe	500
Thể thao	500

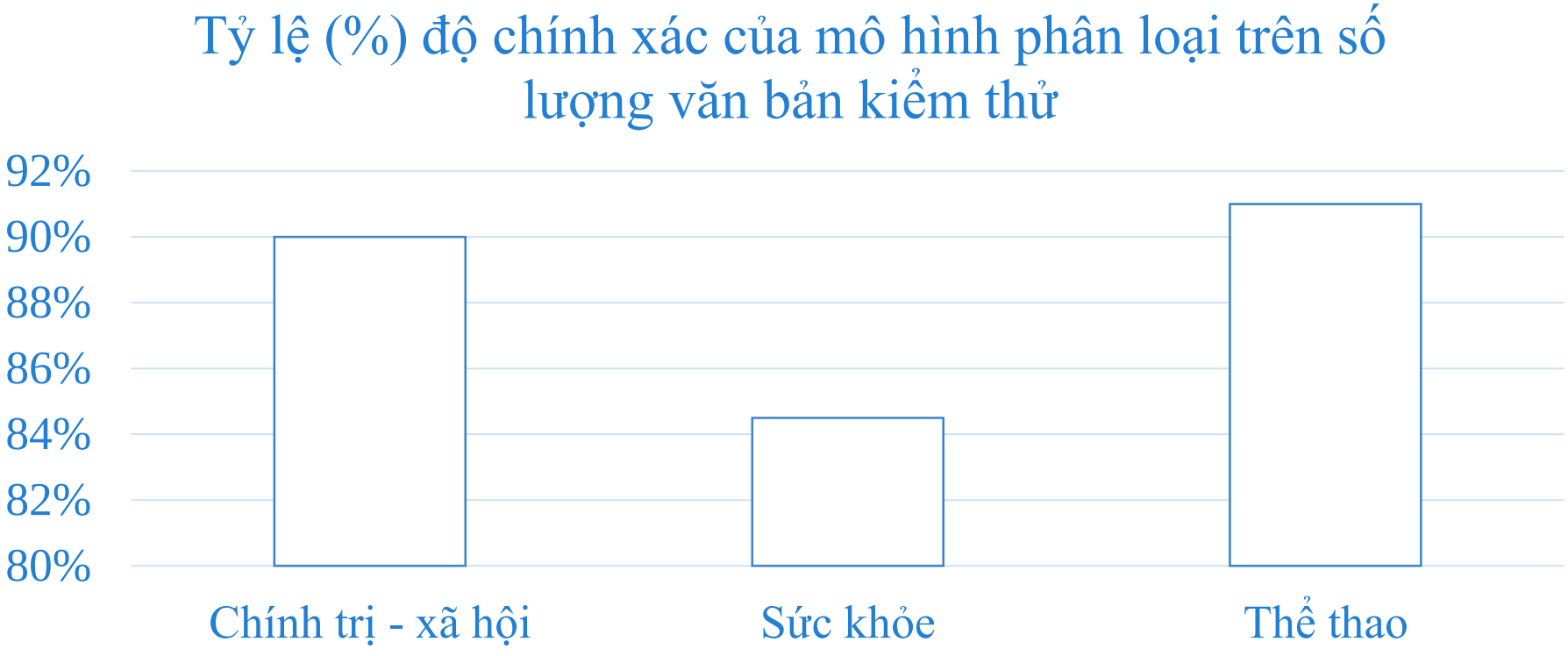
KẾT QUẢ THỰC NGHIỆM

Tên chủ đề	Độ chính xác (Precision)	Độ phủ (Recall)	Độ đo F1 (F-measure)
Chính trị - xã hội	0.87	0.9	0.885
Sức khỏe	0.955	0.845	0.897
Thể thao	0.843	0.91	0.875
Trung bình	0.889	0.885	0.886

KẾT QUẢ THỰC NGHIỆM

Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao
Chính trị - xã hội	180	5	15
Sức khỏe	12	169	19
Thể thao	15	3	182

KẾT QUẢ THỰC NGHIỆM



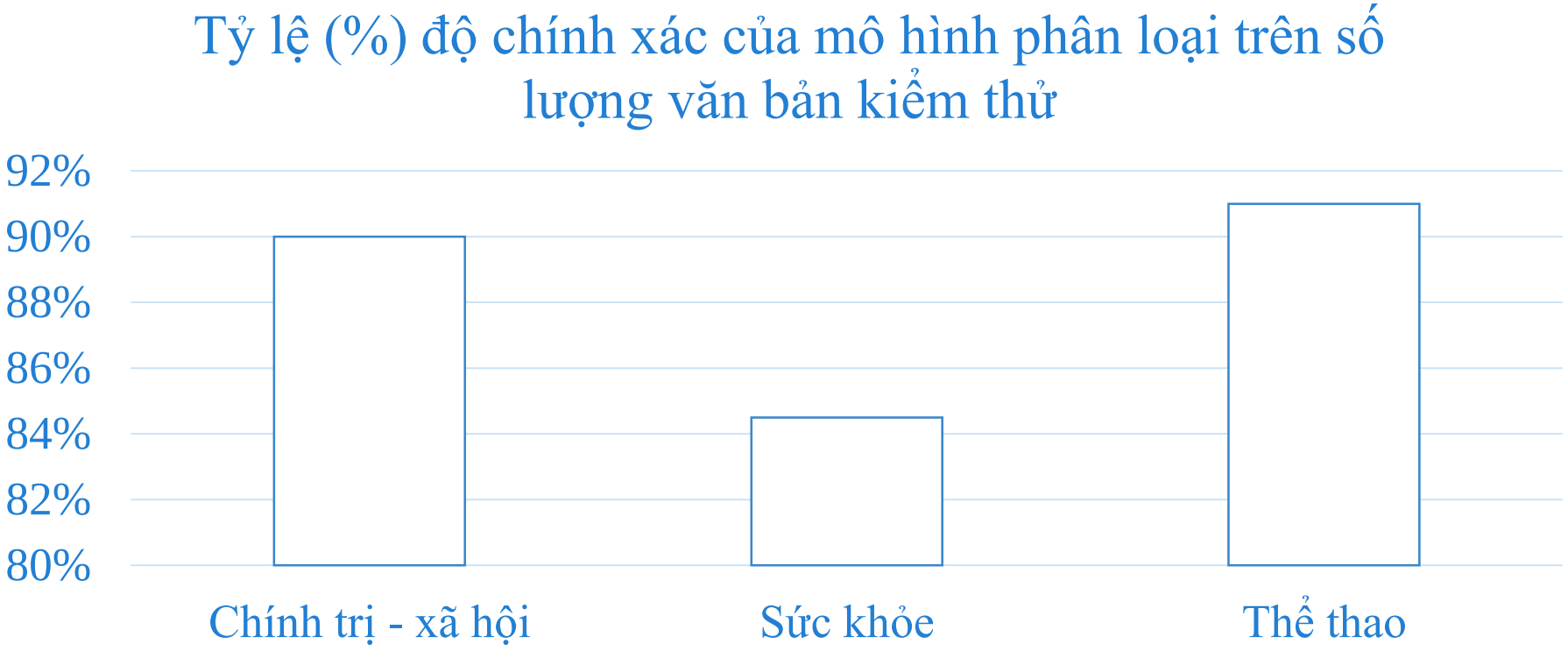
- Tập dữ liệu huấn luyện: 500 văn bản, số tài liệu kiểm tra: 200 văn bản

THỰC NGHIỆM VÀ ĐÁNH GIÁ

Thực nghiệm phân lớp khi gộp chung các văn bản thuộc ba chủ đề:

Tên chủ đề	Chính trị - xã hội	Sức khỏe	Thể thao	Số lượng
Chính trị - xã hội	180	5	15	600
Sức khỏe	12	169	19	
Thể thao	15	3	182	

KẾT QUẢ THỰC NGHIỆM



- Tập dữ liệu huấn luyện: 500 văn bản, số tài liệu kiểm tra: 200 văn bản

KẾT LUẬN & HƯỚNG PHÁT TRIỂN

Đóng góp của luận văn

- Tính TF - IDF nhằm loại bỏ các hư từ trong tiếng Việt, bằng cách loại bỏ những từ có trọng số thấp hơn ngưỡng trung bình. Làm giảm số lượng đồ thị con phổ biến nên số chiều của vec tơ văn bản cũng giảm theo. Từ đó việc huấn luyện và phân lớp sẽ nhanh và chính xác hơn.
- Đóng góp một hướng tiếp cận mới cho bài toán phân loại văn bản tiếng Việt.
- Xây dựng mô hình phân loại văn bản dựa trên mô hình đồ thị có thể ứng dụng trong thực tế.

HẠN CHẾ & KHÓ KHĂN

- Chưa xây dựng được nhiều bộ dữ liệu dùng để huấn luyện và đánh giá.
- So sánh mô hình xây dựng với các mô hình khác như: phân loại văn bản dựa vào mô hình không gian vectơ truyền thống, sử dụng cây quyết định, …

HƯỚNG PHÁT TRIỂN

- Làm giàu thêm bộ dữ liệu huấn luyện để nâng cao độ chính xác trong phân lớp văn bản.
- Xây dựng và đánh giá mô hình với nhiều bộ dữ liệu thực tế hơn nữa.
- Đưa mô hình đồ thị có hướng vào trong bài toán xây dựng đồ thị văn bản.
- Áp dụng mô hình đề xuất cho các bài toán thực tế khác như: khai phá các mạng xã hội, phân loại email, ...

Cảm ơn Quý Thầy Cô và các bạn đã lắng nghe