Trường Đại học Bách Khoa Thành phố Hồ Chí Minh

Khoa Khoa học và Kỹ thuật máy tính

PHÂN TÍCH CẢM XÚC TRONG VĂN BẢN Y KHOA

Giáo viên hướng dẫn:
GS. TS. Cao Hoàng Trụ

Giáo viên phản biện:
GS. TS. Phan Thị Tươi

Sinh viên thực hiện:

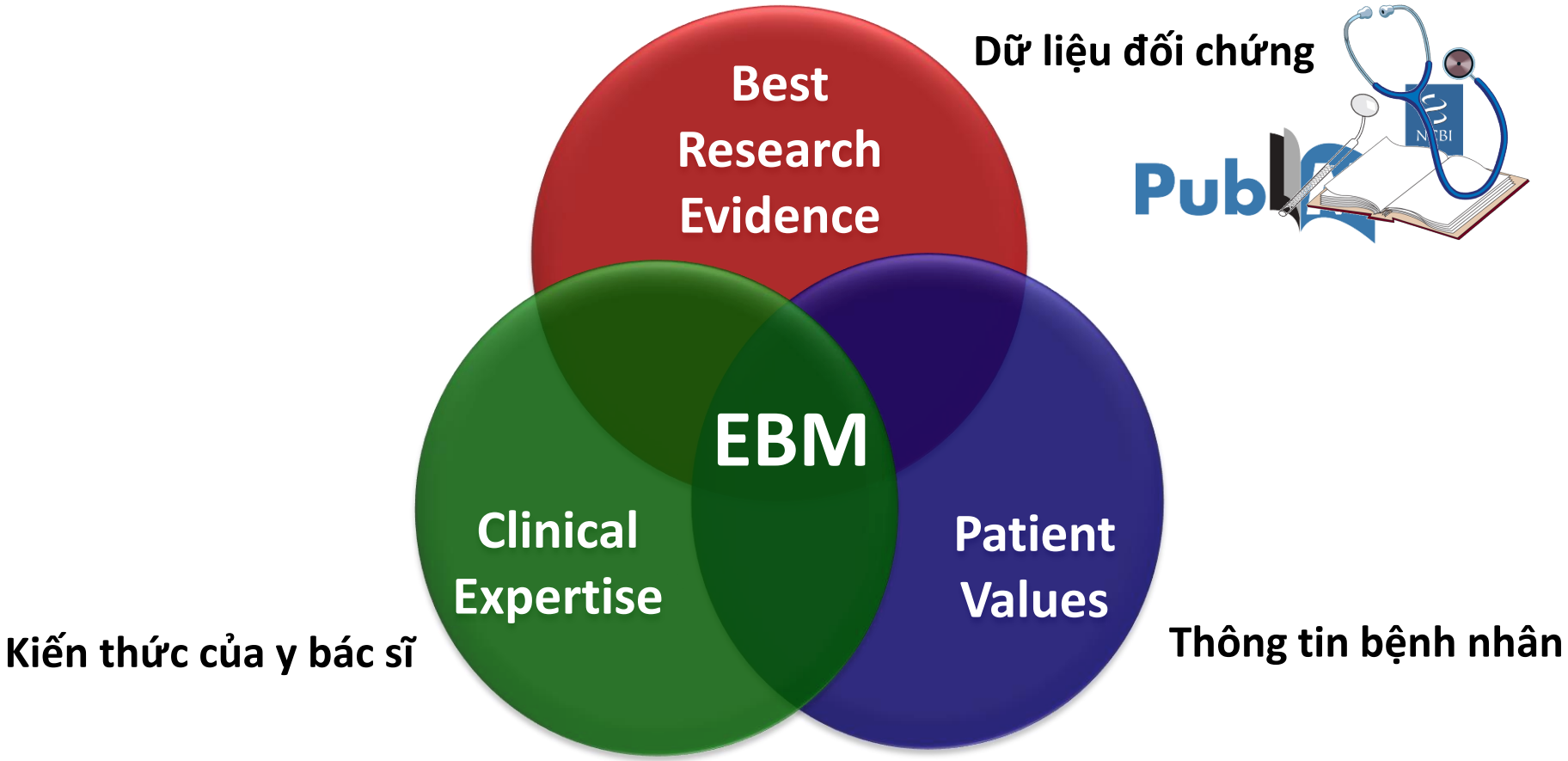
Nguyễn Đức Trí
Nguyễn Diệp Phương Linh

TP. HCM, 01/2017

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

Giới thiệu đề tài



Mô hình Y học thực chứng (EBM)

Sackett D et al, “Evidence-Based Medicine”, 2000.

Giới thiệu đề tài

An evaluation of a chemical cautery agent and an anti-inflammatory ointment for the treatment of recurrent aphthous stomatitis: a pilot study.

Rhodus NL1, Bereuter J.

Abstract

OBJECTIVE: Recurrent aphthous stomatitis is a very common condition, currently treated with anti-inflammatory agents, which palliate the symptoms. The purpose of this clinical trial was to compare a medication commonly used to treat recurrent aphthous stomatitis, Kenalog-in-Orabase, and a newer agent, Debacterol.

METHOD AND MATERIALS: Sixty patients diagnosed with recurrent aphthous stomatitis were enrolled in the study. Twenty patients were assigned to each of the two treatment groups, and 20 age- and sex-matched patients were assigned to the control group, which received no treatment....

RESULTS: In both treatment groups, by day 10, 100% of the ulcers had clinically healed and were no longer causing pain. Patients in the Debacterol group reported a significantly greater decrease in pain at 3 days (> 70%) than did subjects in the other groups (< 20%), although the size of the ulcer did not differ significantly in any of the groups....

CONCLUSION: Patients subjectively reported significantly greater relief from symptoms with Debacterol than with Kenalog-in-Orabase or no treatment. The relief of symptoms associated with recurrent aphthous stomatitis may or may not correspond to clinical improvement, and these two topical medications may affect signs and symptoms of the lesions differently.

Giới thiệu đề tài

CỰC CẢM XÚC	ĐỊNH NGHĨA	VÍ DỤ
TÍCH CỰC	• Kết quả tốt hơn • Cải thiện rõ rệt so với tổng thể dù vẫn có tác dụng phụ	*Patients reported significantly greater relief from symptoms with Debacterol than with Kenalog.*
TIÊU CỰC	• Kết quả xấu hơn • Không đem lại hiệu quả	*We found this expensive therapy to be much less effective than previously believed.*
TRUNG TÍNH	• Không thể hiện kết quả • Đồng thời có nhiều ý kiến tốt xấu mà không có sự lấn át rõ ràng	*We investigated a heterogeneous group of male and female patients.*

Ý nghĩa:

- ✓ Trả lời câu hỏi về tác động của một can thiệp y tế
- ✓ Giúp người điều trị có cái nhìn tổng quát hơn

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

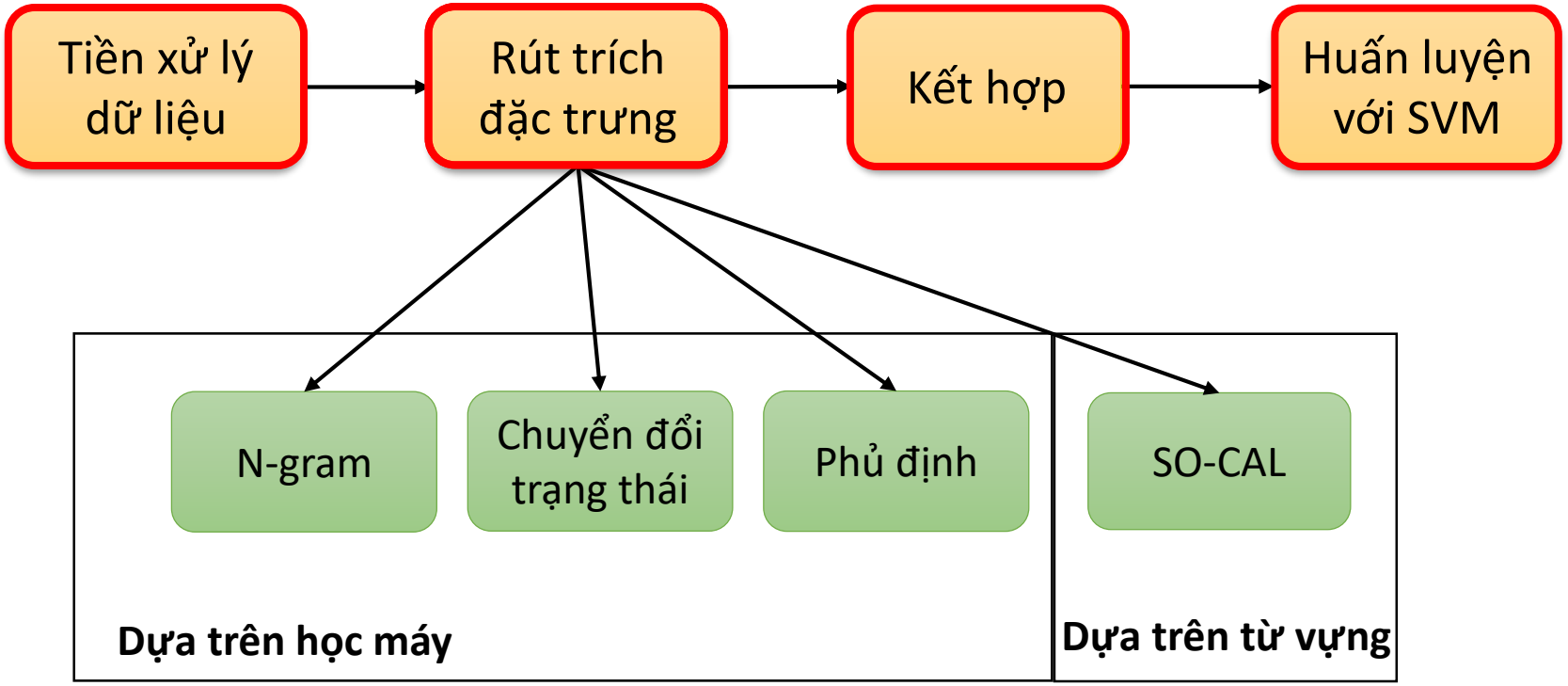
Công trình liên quan

STT	Tác giả	Lĩnh vực phân tích	Phương pháp	Chi tiết
1	Bo Pang, Lillian Lee, Shivakumar Vaithyanathan	Bình luận phim	Dựa trên học máy: Naïve Bayes, Maximum Entropy, SVM	Uni-gram, Bi-gram, POS tagging
2	Lei Xia, Anna Lisa Gentile, James Munro, José Iria	Ý kiến bệnh nhân	Dựa trên học máy: Multinomial Naïve Bayes	Bag of word
3	Yun Niu, MSc, Xiaodan Zhu, MSc, Jianhua Li, MSc, Graeme Hirst	Văn bản nghiên cứu y khoa	Dựa trên học máy: SVM	Uni-gram, Bi-gram, Chuyển đổi trạng thái (Change phrase), Phủ định, Phạm trù (Category)
4	Bruno Ohana, Brendan Tierney	Bình luận phim	Dựa trên từ vựng	Sử dụng từ điển SentiWordNet
5	Maite Taboada, Julian Brooke, Milan Tofilosk, Kimberly Voll, Manfred Stede	Các loại bình luận khác nhau: trò chơi, phim, blog.	Dựa trên từ vựng	Dựa trên các từ điển General Inquirer, WordNet và một số nguồn dữ liệu khác
6	Jin-Cheon Na, Wai Yan Min	Bình luận thuốc	Dựa trên từ vựng	Dựa trên các từ điển Subjective Lexicon, SentiWordNet và một số từ điển khác
7	Abeed Sarker, Diego Mollá-Aliod, Cécile Paris	Các đoạn văn trong báo cáo khoa học thuộc lĩnh vực y khoa	Kết hợp dựa trên học máy và dựa trên từ vựng	N-gram (sử dụng các giải thuật Naïve Bayes, BayesNet, C4.5); từ điển General Inquirer

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

Quy trình huấn luyện



Mô hình huấn luyện bộ phân loại

Rút trích đặc trưng

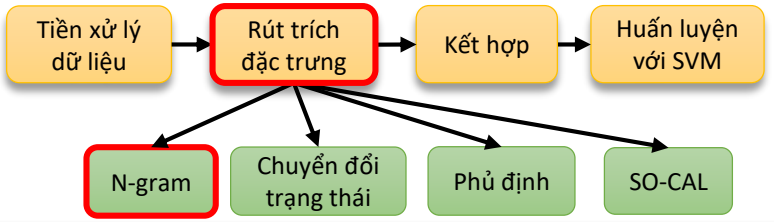
ĐẶC TRƯNG N-GRAM

- Các bước trích xuất:
  - Bước 1: Xây dựng tập từ vựng T
  - Bước 2: Ánh xạ 1 câu sang 1 *vector v* có *n* chiều với n=|T|

Câu 1: *"In addition, it can **avoid** possible side-effect of **medication**."*

Câu 2: *"Using this **medication** did not **result** in patient recuperation."*

Tập từ vựng T	**…**	**avoid**	**result**	**medication**	**…**
Câu 1	…	1	0	1	…
Câu 2	…	0	1	1	…

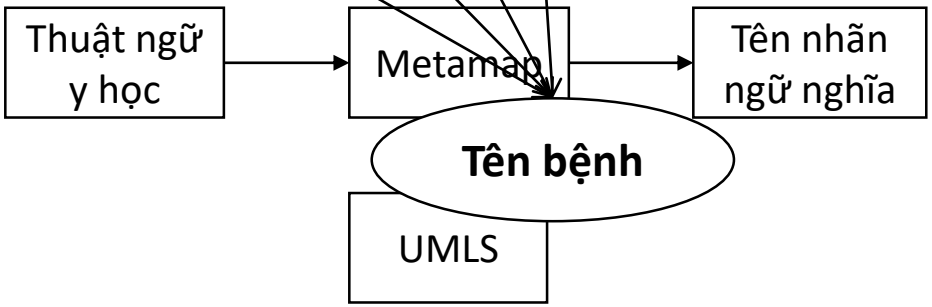


Rút trích đặc trưng

ĐẶC TRƯNG N-GRAM KẾT HỢP METAMAP

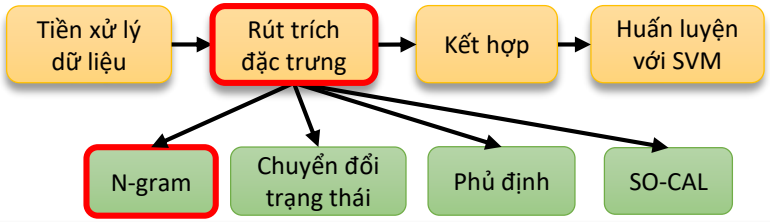
*"Studies are needed to clarify the risk of **stroke** among users who may be susceptible on the basis of age, smoking, **obesity**, **hypertension**, or **migraine** history."*

"Studies are needed to clarify the risk of **DSYN** *among users who may be susceptible on the basis of age, smoking,* **DSYN** , **DSYN** , *or* **DSYN** *history."*



Sử dụng Metamap để gán nhãn ngữ nghĩa

- UMLS (Unified Medical Language System)
- Metamap: gán nhãn ngữ nghĩa



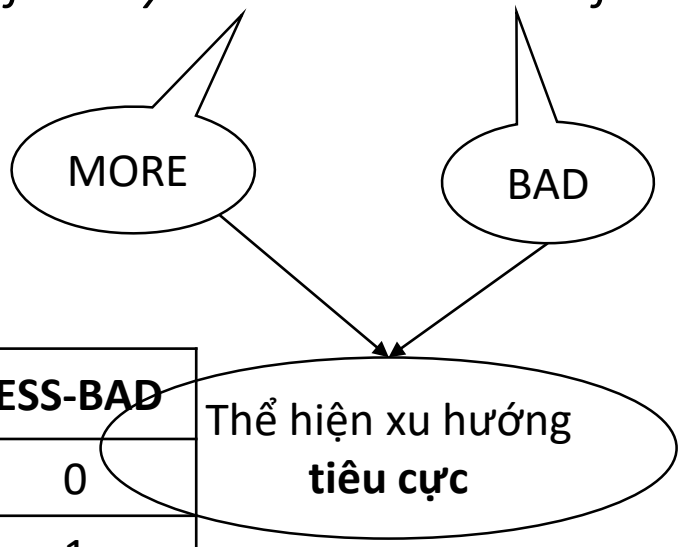
Rút trích đặc trưng

ĐẶC TRƯNG CHUYỂN ĐỔI TRẠNG THÁI (Change phrase)

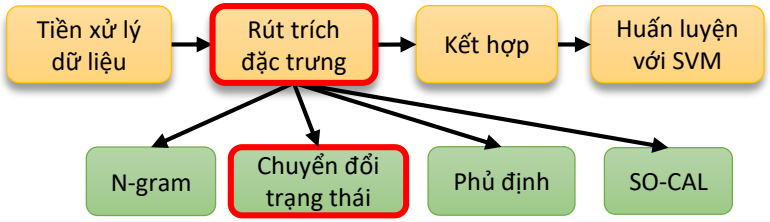
Câu 1: *"Migraine in women of childbearing age significantly **increases** the **risk** of ischaemic but not haemorrhagic stroke."*

Nhận dạng sự chuyển đổi trạng thái thông qua:

- Sự thay đổi trạng thái: LESS và MORE
- Tính chất của trạng thái: GOOD và BAD



	MORE-GOOD	MORE-BAD	LESS-GOOD	LESS-BAD
Câu 1	0	1	0	0
Câu 2	0	0	0	1
…	…	…	…	…

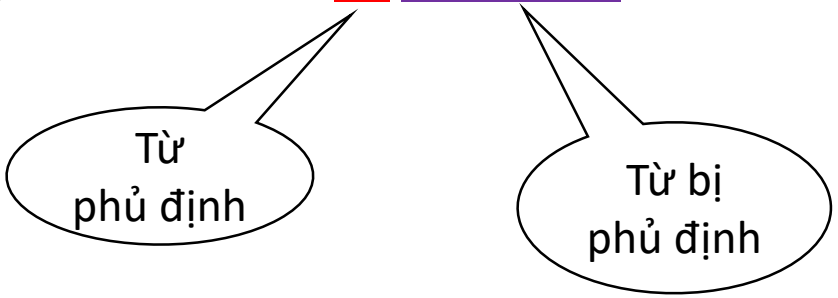


Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of Polarity Information in Medical Text," *Proc. Am. Med. Informatics Assoc. Symp.*, pp. 570–574, 2005.

Rút trích đặc trưng

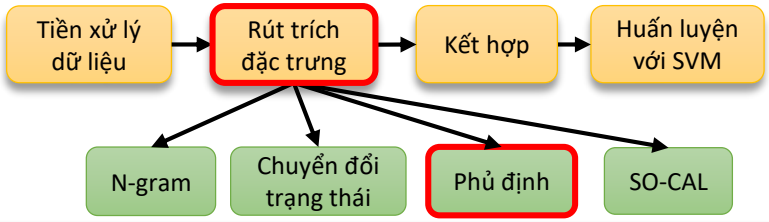
ĐẶC TRƯNG PHỦ ĐỊNH

"Patients subjectively reported significantly greater relief from symptoms with Debacterol than with Kenalog-in-Orabase or no treatment."



Giải thuật xử lý phủ định NegEx:

- Xác định từ phủ định
- Xác định tầm vực phủ định

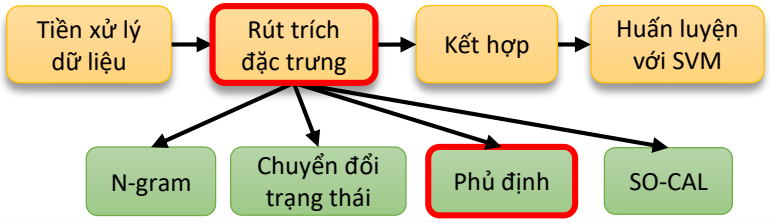


H. Tanushi, H. Dalianis, M. Duneld, M. Kvist, M. Skeppstedt, and S. Velupillai, "Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg," 2013.

Rút trích đặc trưng

ĐẶC TRƯNG PHỦ ĐỊNH

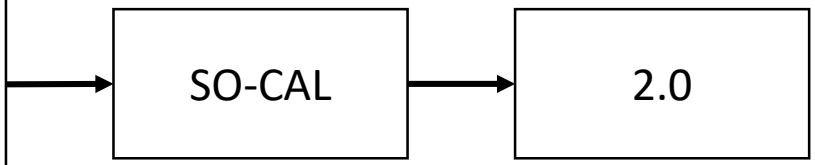
- Cách 1: Gắn nhãn `_NEG` cho từ bị phủ định
- Cách 2: Gắn nhãn `NEGATION` cho từ phủ định



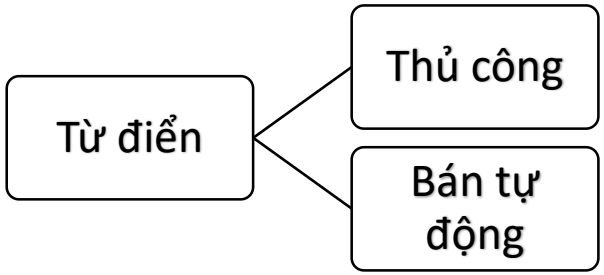
Rút trích đặc trưng

ĐẶC TRƯNG SO-CAL (Semantic Orientation CALculator)

"Patients reported significantly greater relief from symptoms with Debacterol than with Kenalog-in-Orabase."

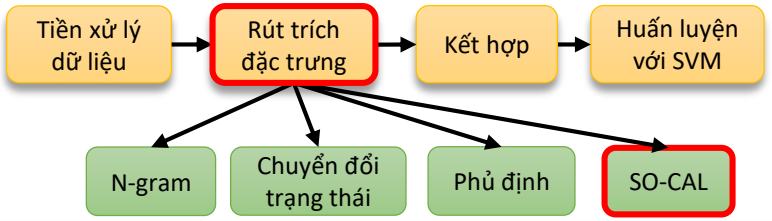


$$\text{Điểm số}_{\text{câu}} = f(\text{Điểm số}_{\text{từ}})$$



Từ	Giá trị
monstrosity	-5
hate (noun and verb)	-4
disgust	-3
sham	-3
fabricate	-2
delay (noun and verb)	-1
determination	1

Một số từ trong từ điển



M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Assoc. Comput. Linguist.*, pp. 267–307, 2011.

Rút trích đặc trưng

ĐẶC TRƯNG SO-CAL

$$Điểm\ số_{câu} = f(Điểm\ số_{từ})$$

Hàm f : Điểm số của câu bằng trung bình cộng điểm số của các từ có giá trị khác 0.

1. Từ loại: tính từ, động từ, danh từ, trạng từ

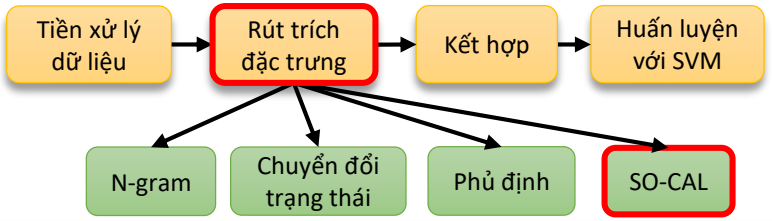
   Từ "***novel***":

   - Danh từ: trung tính → 0.0
   - Tính từ: tích cực → 1.0

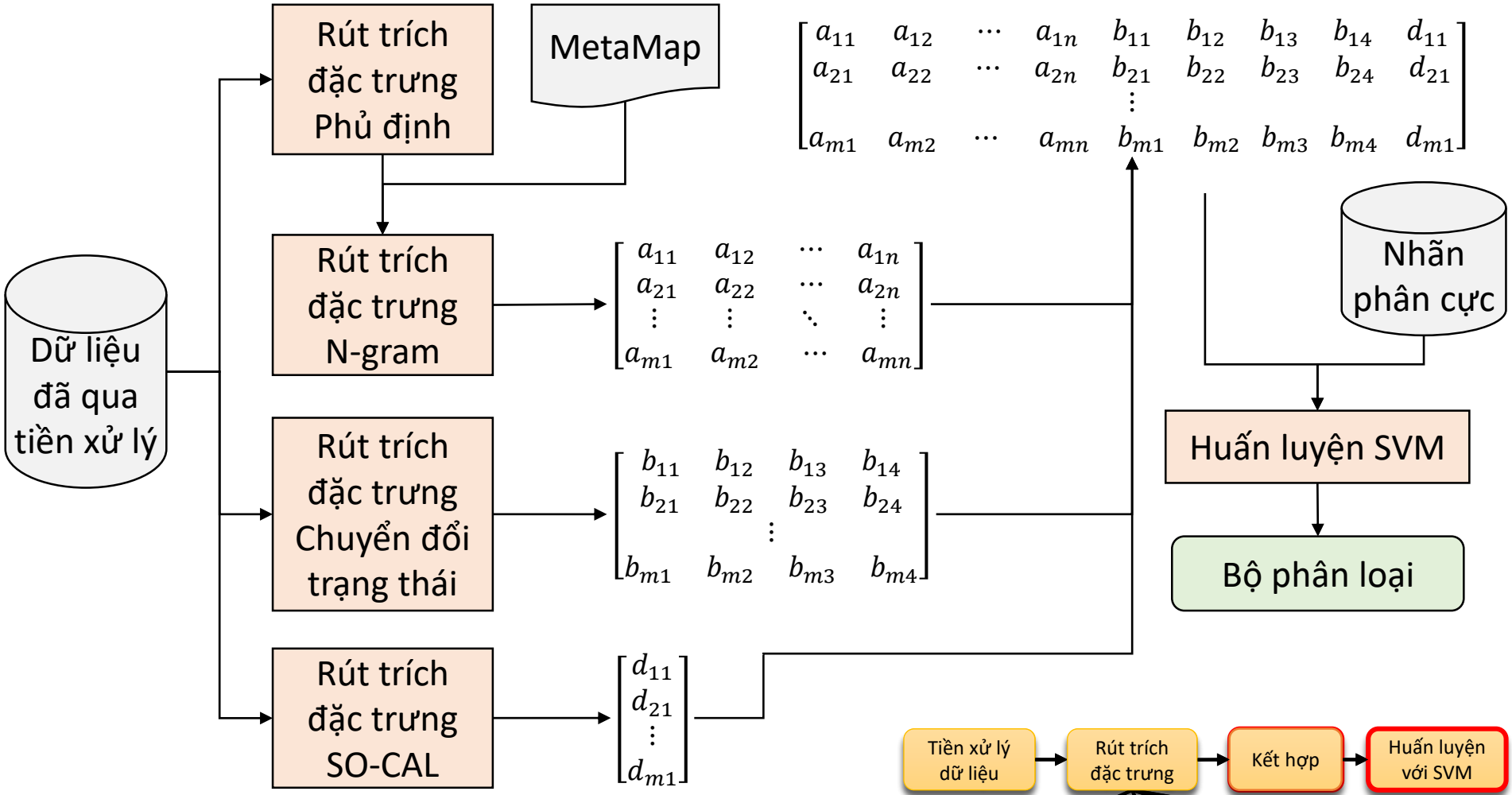
2. Tính tăng cường:

$$Điểm\ số_{cụm\ từ} = Điểm\ số_{từ\ bị\ tác\ động} \times (100\% + tỉ\ lệ\ tác\ động) \quad (1)$$

"very good" → $\mathbf{3.0} \times (100\% + \mathbf{25\%}) = \mathbf{3.75}$



Kết hợp các đặc trưng

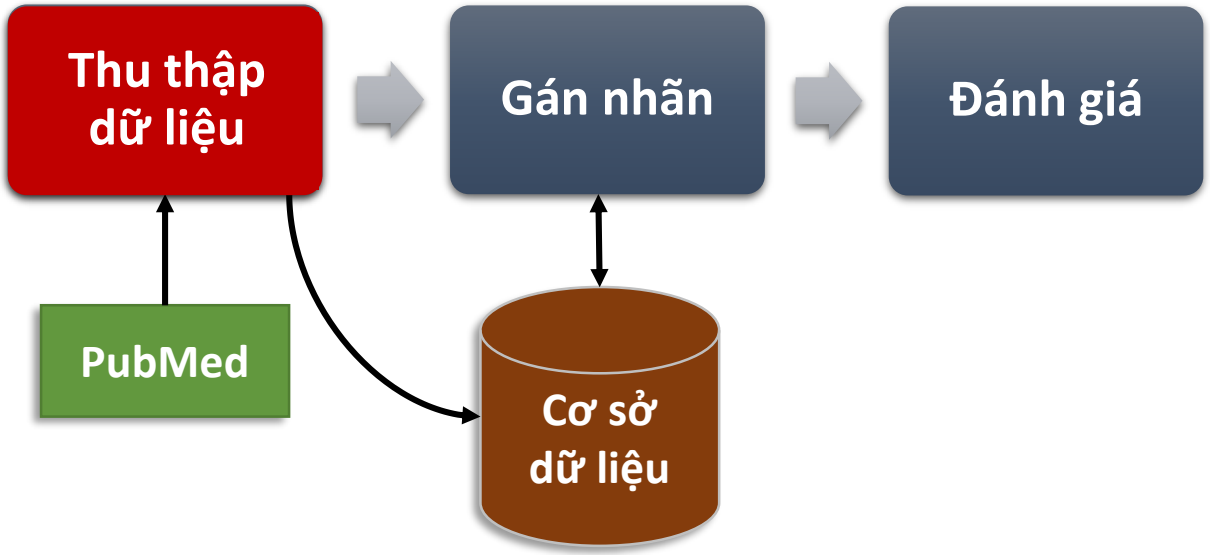


Mô hình kết hợp các đặc trưng

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

Xây dựng tập dữ liệu

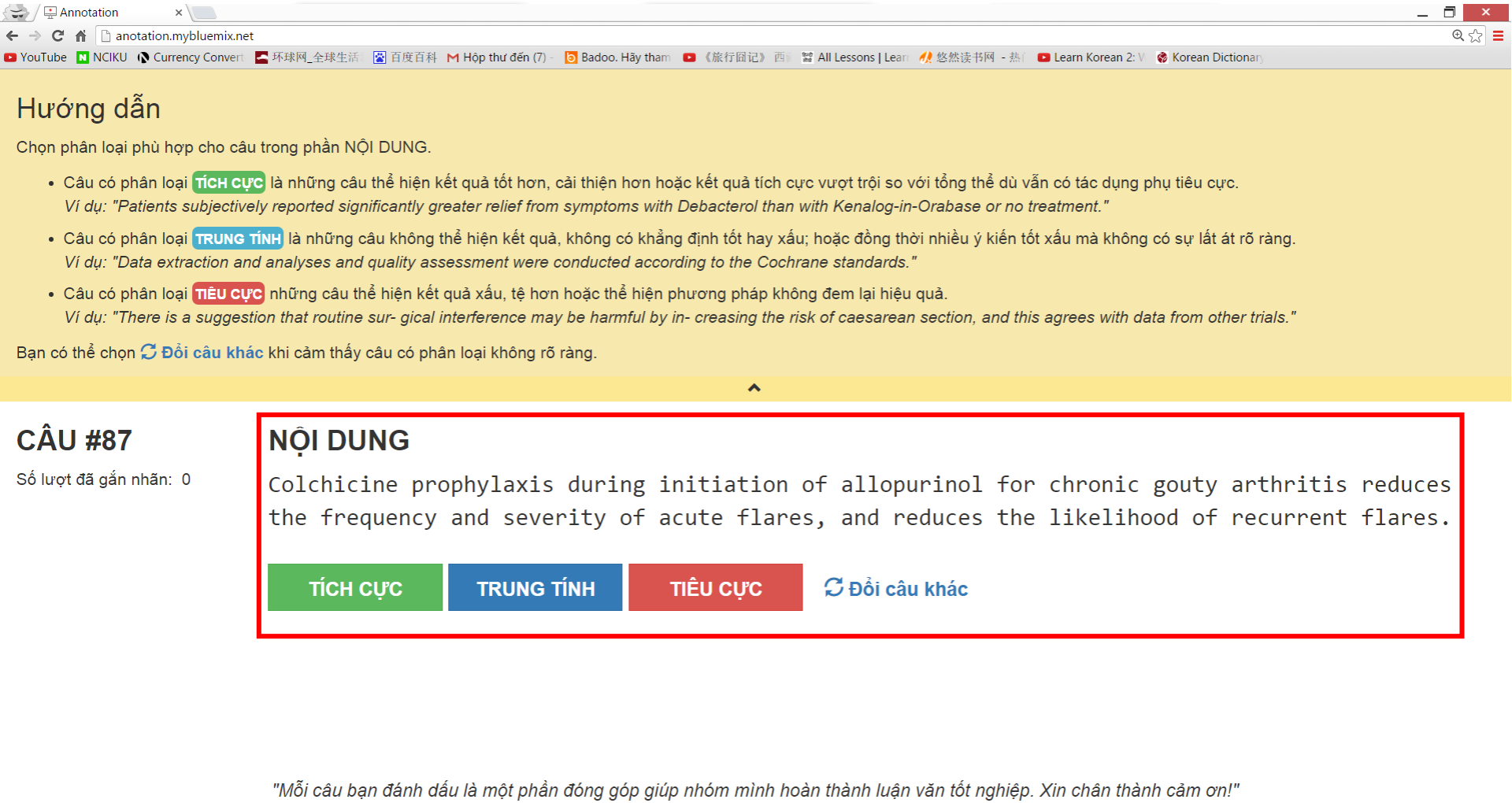


Quy trình xây dựng tập dữ liệu

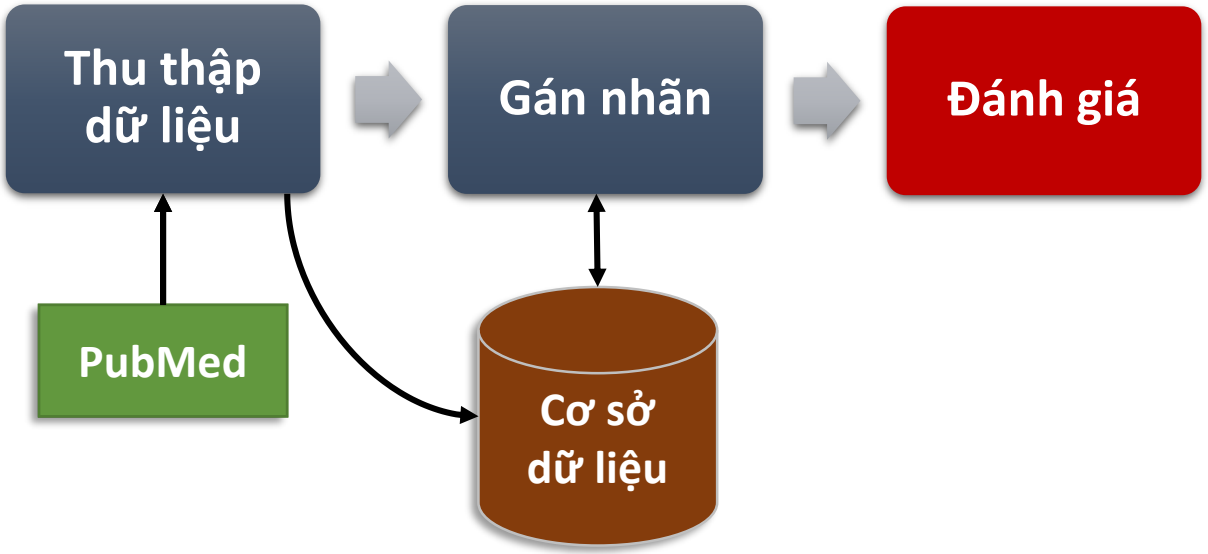
Id	Câu	Cực cảm xúc
191	Patients subjectively reported significantly greater relief from symptoms with Debacterol than with Kenalog-in-Orabase or no treatment.	
122	The relief of symptoms associated with recurrent aphthous stomatitis may or may not correspond to clinical improvement, and these two topical medications may affect signs and symptoms of the lesions differently.	

Mẫu dữ liệu thu thập được

Xây dựng tập dữ liệu



Xây dựng tập dữ liệu



Quy trình xây dựng tập dữ liệu

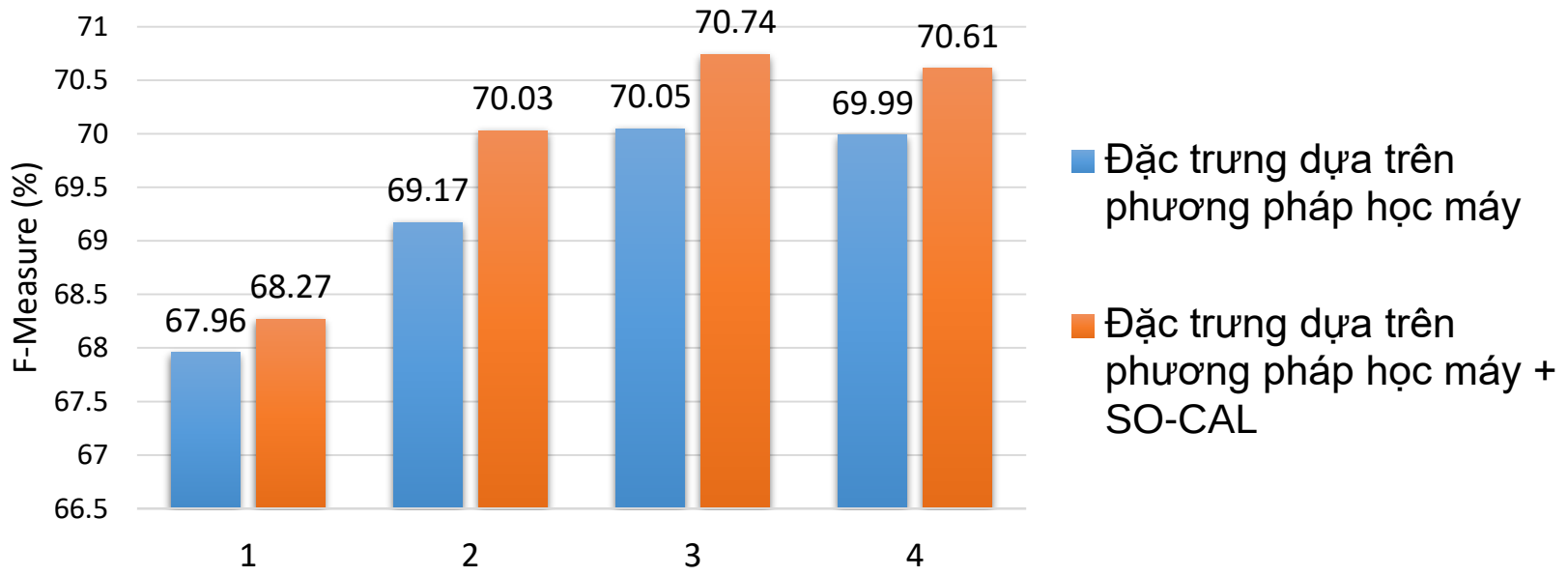
- Số lượng người gán nhãn: 16 sinh viên
- Tổng số câu được gán nhãn là 552 câu
- Hệ số Fleiss's kappa đạt khá tốt ($\kappa = 72.54\%$)

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

Kết quả thí nghiệm

Kết hợp các đặc trưng dựa trên phương pháp học máy với đặc trưng SO-CAL



Thí nghiệm	Đặc trưng dựa trên phương pháp học máy	Kết quả (%)	Kết quả kết hợp đặc trưng SO-CAL (%)
1	N-gram	67.96	68.27
2	N-gram + Chuyển đổi trạng thái	69.17	70.03
3	N-gram + Chuyển đổi trạng thái + Phủ định	70.05	**70.74**
4	N-gram kết hợp MetaMap + Chuyển đổi trạng thái + Phủ định	69.99	70.61

Nội dung

- Giới thiệu đề tài
- Công trình liên quan
- Phương pháp đề xuất
- Xây dựng tập dữ liệu
- Kết quả thí nghiệm
- Tổng kết

Tổng kết

- Kết quả đạt được:
  - Xây dựng hệ thống phân tích tính phân cực cảm xúc trong văn bản y khoa trên mức câu
  - Xây dựng trang web đánh nhãn dữ liệu, tập dữ liệu thu được có độ đồng nhất khá tốt
- Hạn chế và hướng cải tiến:
  - Tăng kích thước tập dữ liệu
  - Phân tích tập câu có nhãn TIÊU CỰC
- Hướng phát triển:
  - Phát triển hệ thống phân tích tính phân cực cảm xúc trên văn bản y khoa tiếng Việt

Cảm ơn Hội đồng đã lắng nghe