BÀI 6: KHAI PHÁ QUAN ĐIỂM (TIẾP)

Các bài toán thành phần

- 1. Trích rút khía cạnh
  - "*The voice quality of this phone is amazing*"
  - "*I love this phone*" (khía cạnh GENERAL)
- 2. Phân loại cảm xúc mức khía cạnh
  - "*The voice quality of this phone is amazing*" → TÍCH CỰC
  - "*I love this phone*" → TÍCH CỰC

Phân loại cảm xúc mức khía cạnh

- Hướng tiếp cận học có giám sát
  - Dựa trên cú pháp phụ thuộc để trích rút đặc trưng cú pháp
  - Đạt kết quả cao nhưng khó điều chỉnh vào lĩnh vực mới
- Hướng tiếp cận dựa trên từ điển
  - Đạt kết quả cao trên nhiều lĩnh vực
  - Tuy nhiên cần có hiểu biết về ngôn ngữ và lĩnh vực, sử dụng nhiều luật

Nội dung

[1] Sử dụng cây tri thức cảm xúc

[2] Phân tích cảm xúc trên twitter

[3] Phân tích cảm xúc trên mạng xã hội

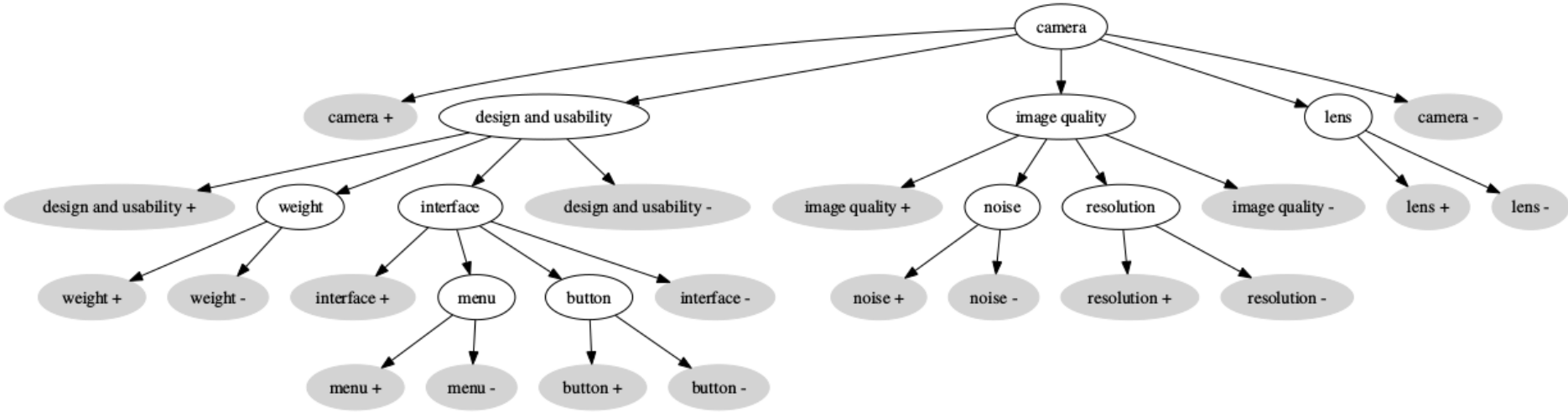
[4] Phát hiện & gán thực thể

[5] Khai phá câu so sánh

[1] Sử dụng cây tri thức cảm xúc

- Phân loại phân cấp dựa trên kĩ thuật học phân cấp
- Cây tri thức cảm xúc (SOT):
  - Thể hiện mối liên hệ cha con giữa các khía cạnh trong miền
  - Mỗi khía cạnh đi kèm với các nút thể hiện cảm xúc đối với khía cạnh đó

Minh họa cây tri thức cảm xúc



SOT

- $T(v, v^+, v^-, \mathbf{T})$
- v: nút gốc thể hiện thuộc tính v
- v^+: nút tích cực ứng với thuộc tính v
- v^-: nút tiêu cực ứng với thuộc tính v
- **T**: tập các cây SOT con của T: $T'(v', v'^+, v'^-, \boldsymbol{T'})$

HL-SOT

- Câu $x \in X$, $X = R^d$
- Tập các nút trong cây: $Y = \{1, 2, \ldots, N\}$
- Véc-tơ nhãn của x: $y = \{y_1, y_2, \ldots, y_N\} \in \{0,1\}^N$
- $\forall i \in Y$,
  - $y_i = 1$ nếu x được gán nhãn bởi bộ phân loại của nút i
  - $y_i = 0$ nếu x không được gán nhãn bởi bộ phân loại của nút i

Phát biểu bài toán

- $y \in \{0,1\}^N$ *đáp ứng một cây SOT khi và chỉ khi*
  - $\forall i \in Y, \forall j \in A(i)$: *nếu* $y_i = 1$ *thì* $y_j = 1$, *trong đó* $A(i)$ *là tập các nút tổ tiên của* i
- Gọi tập hợp các véc-tơ nhãn đáp ứng SOT là τ
- Học mô hình phân loại phân cấp $f: X \rightarrow \tau$ để sinh ra véc-tơ y cho mỗi văn bản đầu vào x sao cho y thỏa mãn SOT

HL-SOT

- $y = f(x) = g(W \cdot x)$
- $W = (w_1, ..., w_N)^T$
  - w_i là trọng số của bộ phân loại tuyến tính của nút I
- $y_i = w_i^T x \geq \theta_i$ nếu i là nút gốc hoặc $y_j = 1$ với $\forall j \in A(i)$; ngược lại $y_i = 0$
  - θ_i *là* ngưỡng của bộ phân loại của nút i

Học tham số

- Cho tập DL huấn luyện $D = \{(r, l) \mid r \in X, l \in Y\}$
- Ma trận trọng số W được khởi tạo = 0
- Véc-tơ ngưỡng θ được khởi tạo = 0

Học tham số (2)

- Với một ví dụ r_t, trọng số được cập nhật như sau:

$$w_{i,t} = (I + S_{i,Q(i,t-1)} S_{i,Q(i,t-1)}^{\top} + r_t r_t^{\top})^{-1} \times S_{i,Q(i,t-1)} (l_{i,i_1}, l_{i,i_2}, \ldots, l_{i,i_{Q(i,t-1)}})^{\top}$$

- I_{dxd} : ma trận định danh
- $Q(i, t\text{-}1)$: số lần cha của nút i được gán positive trước đó
- $S_{i,Q(i,t\text{-}1)} = [r_{i,1}, \ldots, r_{i,Q(i,t\text{-}1)}]$
- Chỉ cập nhật trọng số $w_{i,t}$ của những nút i có nút cha được gán positive

Học tham số (3)

- Cập nhật ngưỡng của bộ phân loại

$$\theta_{t+1} = \theta_t + \epsilon(\hat{y}_{r_t} - l_{r_t}),$$

- trong đó ε là một sô dương nhỏ đê điều chỉnh tốc độ cập nhật
- Bộ phân loại dự đoán đúng, không cần cập nhật
- Nếu gán nhầm thuộc tính là positive, cần tăng ngưỡng θ
- Nếu bỏ sót thuộc tính (gán negative), cần giảm ngưỡng θ

Giải thuật học tham số

Algorithm 1 Hierarchical Learning Algorithm HL-SOT

INITIALIZATION:

Each vector $w_{i,1}, i = 1, ..., N$ of weight matrix W_1 is set to be 0 vector

Threshold vector θ_1 is set to be 0 vector

BEGIN

for $t = 1, ..., |D|$ **do**

    Observe instance $r_t \in \mathcal{X}$

    for $i = 1, ...N$ **do**

        Update each row $w_{i,t}$ of weight matrix W_t by Formula 1

    end for

    Compute $\hat{y}_{r_t} = f(r_t) = g(W_t \cdot r_t)$

    Observe label vector $l_{r_t} \in \mathcal{Y}$ of the instance r_t

    Update threshold vector θ_t by Formula 2

end for

END

[2] Phân tích cảm xúc trên twitter

- Tweet chứa tối đa 140 kí tự
- 2011: Twitter có 190M người dùng, mỗi ngày có 65M tweet
- Người dùng có xu hướng bày tỏ cảm xúc trên Twitter
- Một số công cụ phân tích cảm xúc trên Twitter: Tweetfeel, Twendz, Twitter Sentiment

Tính chất các tweet

- Tweet thường ngắn và nhập nhằng hơn so với bình luận sản phẩm
- Bình luận thường đã biết đối tượng được đánh giá; trong khi đó cần xác định đối tượng được đánh giá trong tweet
- Các tweet liên quan cung cấp thêm ngữ cảnh cho bộ phân loại
- Các phương pháp phân loại không phụ thuộc đối tượng không phù hợp với phân loại tweet

Phát biểu bài toán

- Đầu vào: Một tập các tweet chứa đối tượng cần đánh giá
- Đầu ra: Phân loại cảm xúc của mỗi tweet đối với đối tượng
  - Trung tính: Không thể hiện cảm xúc
  - Tích cực
  - Tiêu cực

Các bước của thuật toán

1. Phân loại chủ quan/khách quan: Nếu tweet được phân loại khách quan → thể hiện cảm xúc trung tính
2. Phân loại cảm xúc tích cực và tiêu cực
3. Tối ưu dựa trên độ thị gồm các tweet liên quan

- Sử dụng bộ phân loại *SVM* với nhân tuyến tính (công cụ *SVMLight* với các tùy chọn mặc định)

Tiền xử lý

- Gán nhãn từ loại sử dụng OpenNLP
- Stemming sử dụng từ điển gồm 20,000 mục từ (vd ‘playing’ → ‘play’)
- Chuẩn hóa dựa trên luật đơn giản (vd ‘gooood’ → ‘good’, ‘luve’ → ‘love’)
- Phân tích cú pháp phụ thuộc sử dụng Minimum Spanning Tree

Các đặc trưng độc lập

- Đặc trưng nội dung: từ, dấu câu, emoticon, hashtag
- Đặc trưng từ vựng: Sử dụng từ vựng cảm xúc của General Inquirer
- Đây là các đặc trưng sử dụng phổ biến trong các bộ phân loại cảm xúc không phụ thuộc đối tượng

Tập đối tượng mở rộng

1) Các cụm danh từ

"*I am passionate about Microsoft technologies, especially Silverlight*"

2) Mở rộng dựa trên phân giải đồng tham chiếu

"Oh, Jon Stewart. How I love you so."

3) Top K danh từ và cụm danh từ liên quan nhất với đối tượng dựa trên PMI

Tập đối tượng mở rộng (2)

$$PMI(w,t) = \log \frac{p(w,t)}{p(w)\,p(t)}$$

- p(*w*,*t*): xác suất *w* và *t* cùng x/h trong corpus
- p(*w*): xác suất w x/h trong corpus
- p(*t*): xác suất *t* x/h trong corpus
- K = 20, corpus chứa 20M tweet

Tập đối tượng mở rộng (3)

4) Từ trung tâm của cụm danh từ nếu PMI lớn hơn một ngưỡng cho trước

"Microsoft technologies" → 'technologies'

"the price of iPhone" → 'price'

"LoveGame by Lady Gaga" → 'LoveGame'

Đặc trưng phụ thuộc đối tượng

- Gọi đối tượng là T
- w_i_*arg2*: Ngoại động từ nhận T làm object
  "*I love iPhone*" → '*love_arg2*'
- w_i_*arg1*: Ngoại động từ nhận T làm subject
- w_i_*it_arg2*:Nội động từ nhận T làm subject
- w_i_*arg1*: Danh từ hoặc tính từ nhận T làm từ trung tâm (trong cụm danh từ)

Đặc trưng phụ thuộc đối tượng (2)

- w_i_*cp_arg1*: Danh từ hoặc tính từ được liên kết với T thông qua một copula (động từ "*to be*")
- w_i_*arg*: Tính từ hoặc nội động từ x/h như một câu độc lập và T x/h ở câu trước đó

  "*John did that. Great!*" → '*great_arg*'
- *arg1_v_*w_i: Trạng từ bổ nghĩa cho động từ nhận T làm subject

  "*iPhone works better with the CellBand*" → 'arg1_v_better'

Đặc trưng phụ thuộc đối tượng (3)

- Nếu đặc trưng được bổ nghĩa bởi một từ phủ định thì thêm tiền tố "neg-"

  "*iPhone does not work better with the CellBand*" →
  '*arg1_v_neg-well*', '*neg-work_it_arg1*'

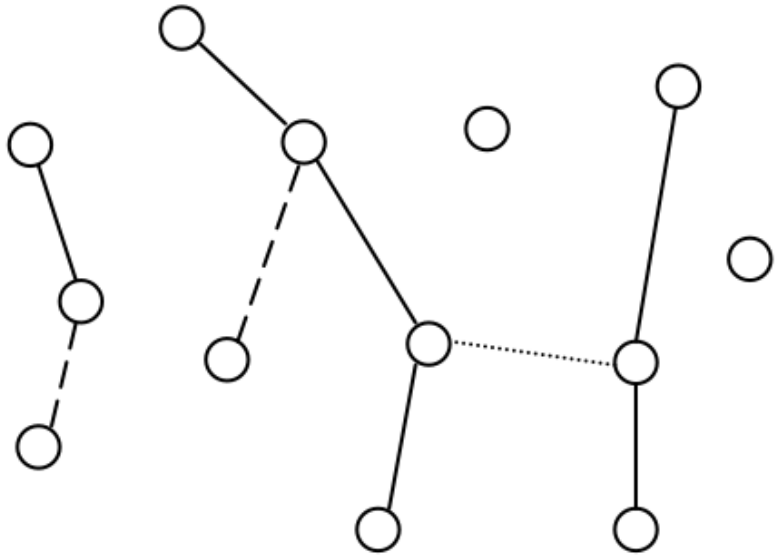
- Tập các từ phủ định: *not, no, never, n't, neither, seldom, hardly*

Tối ưu trên đồ thị

- Chỉ sử dụng nội dung để gán nhãn cảm xúc có thể không chính xác với các tweet ngắn
- Sử dụng thêm các ngữ cảnh sau
  - Retweet: Giữ nguyên nội dung ban đầu
  - Tweet của cùng người dùng có chứa đối tượng trong một khung thời gian ngắn: Giả sử người dùng giữ nguyên cảm xúc với đối tượng
  - Reply: Phản hồi lại tweet hoặc được phản hồi

Đồ thị các tweet

- Nét liền: Tweet của cùng người dùng
- Nét đứt: retweet
- Nét chấm tròn: reply



Đánh giá mô hình

- Tập các truy vấn {*Obama*, *Google*, *iPad*, *Lakers*, *Lady Gaga*}
- Dữ liệu: 459 tích cực, 268 tiêu cực và 1,212 trung tính
- Độ đồng thuận: 86% trên 100 tweet
  - 1 tweet tích cực - tiêu cực
  - 13 tweet trung tính - tích cực/tiêu cực

Đánh giá các đặc trưng

Features	Accuracy (%)
Content features	61.1
+ Sentiment lexicon features	63.8
+ Target-dependent features	**68.2**
Re-implementation of (Barbosa and Feng, 2010)	60.3

Target	Accuracy (%)
Exact target	65.6
+ all extended targets	**68.2**
- co-references	68.0
- targets found by PMI	67.8
- head nouns	67.3

*"No debate needed, heat can't beat **lakers** or celtics"* (negative by TS but positive by human)

*"why am i getting spams from weird people asking me if i want to chat with **lady gaga**"* (positive by TS but neutral by human)

*"Bringing iPhone and **iPad** apps into cars? http://www.speakwithme.com/ will be out soon and alpha is awesome in my car."* (positive by TS but neutral by human)

*"Here's a great article about Monte Veronese cheese. It's in Italian so just put the url into **Google** translate and enjoy http://ow.ly/3oQ77"* (positive by TS but neutral by human)

[3] Phân tích cảm xúc trên MXH

- Phân tích cảm xúc các đánh giá sản phẩm
- Đa ngôn ngữ: English, French, Dutch
- Tập nhãn = {tích cực, tiêu cực, trung tính}
- Sử dụng các đặc trưng ngôn ngữ
- Sử dụng các bộ phân loại dựa trên đặc trưng

Đặc trưng từ vựng

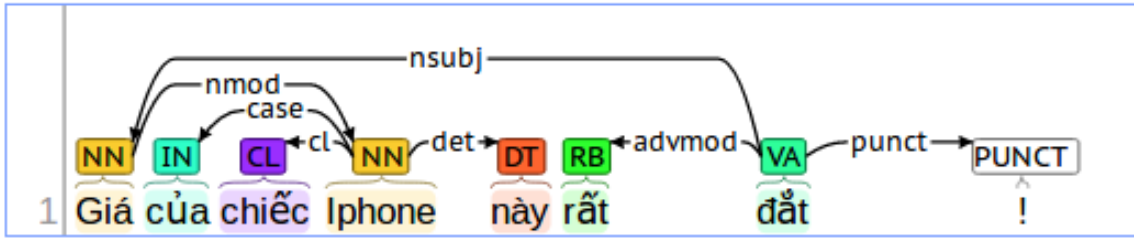
- Unigram: Các từ, token trong câu; loại bỏ các stopword (publist.com)
- Stem: Sử dụng giải thuật Porter (vd 'playing' → 'play')
- Phủ định: vd "*not worth*"
- Đặc trưng văn bản: vd "*mêmе si le film a eu beaucoup de succès, je le trouvais vraiment nul!*" (even though the movie had a lot of success, I really found it nothing!)

Đặc trưng cú pháp

- Chênh lệch độ sâu: Giữa đặc trưng từ và thực thể trong cây cú pháp nghịch đảo với trọng số của từ (English, Dutch)
- Khoảng cách đường đi: Khoảng cách (theo BFS) giữa đặc trưng từ và thực thể trong cây cú pháp nghịch đảo với trọng số của từ (French)
- Khoảng cách: Khoảng cách giữa từ và thực thể nghịch đảo với trọng số của từ

VD đặc trưng cú pháp

- Depth(‘giá’) = 1
- Depth(‘Iphone’) = 2
- Path_distance(‘giá’, ‘Iphone’) = 1



http://45.117.171.213/bknlptool/

Tính khác biệt ngôn ngữ

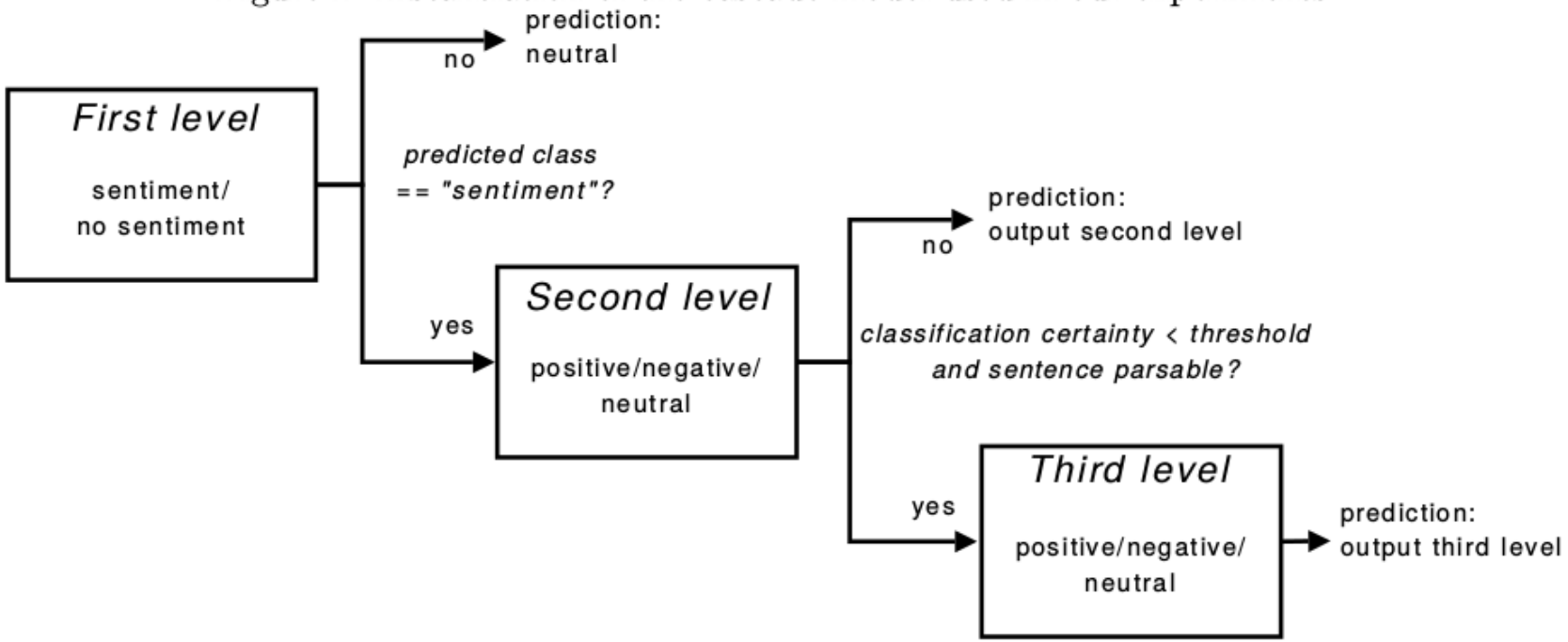
- Compound noun: ‘topfilm’ (top movie)
- Composed verb: “tegenvallen, valt tegen” (to be below expectations), “meevallen, valt mee” (turn out better than expected)
- “je ne suis pas d’accord” (I don’t agree)
- “L é tro bel cet voitur” (Elle est trop belle cette voiture - She is too beautiful, this car)
- j’aime ce film (I love this movie) /ce film est bien (this movie is good)
- de film is goed (this movie is good)

Các mô hình phân loại

- Support Vectors Machines (SVMs): Bộ phân loại học các véc-tơ hỗ trợ để phân loại đối tượng thành hai lớp sao cho biên lớn nhất (Weka)
- Multinomial Naive Bayes (MNB): Bộ phân loại đa lớp dựa trên xác suất Bayes với giả thiết độc lập xác suất giữa các đặc trưng (Weka)
- Maximum Entropy (ME): Bộ phân loại dựa trên cực đại hóa độ hỗn độn (MaxEnt)

Cascade model

Figure 1: Instantiation of the cascade model used in our experiments.



Học chủ động

- Mục tiêu: Học mô hình có chất lượng tốt với số lượng dữ liệu có nhãn tối thiểu
- Phương pháp: Lựa chọn tự động các ví dụ để gán nhãn từ tập không có nhãn (có thể dựa trên một số ví dụ có nhãn gốc)

Lấy mẫu không chắc chắn

- Tiêu chí: Chọn các ví dụ mà bộ phân loại đã có không chắc chắn. Mức độ không chắc chắn dựa trên:
  - Các bộ phân loại xác suất (MNB, ME)
  - Khoảng cách tới siêu phẳng (SVM)
- Mục tiêu:
  - Giảm dư thừa trong dữ liệu huấn luyện
  - Cải thiện các ví dụ nhập nhằng (nhiều cảm xúc, cảm xúc khác nhau đối với các thực thể khác nhau)

Tập dữ liệu

- Bình luận sản phẩm của người dùng và các bài tin tức, đánh giá sản phẩm trên diễn đàn, trang tin
- Blog: skyrock.com, livejournal.com, xanga.com, blogspot.com
- Trang đánh giá: amazon.fr, ciao.fr, kieskeurig.nl
- Diễn đàn tin tức: fok.nl, forums.automotive.com
- Nhiễu: Quảng cáo, spam, phong cách viết riêng

Tập dữ liệu (2)

- Đối tượng đánh giá: ô tô, phim; thay thế tên bằng nhãn chung ‘CAR’ hoặc ‘MOVIE’
- Loại bỏ các câu hỏi
- Phân loại cảm xúc mức câu
- Độ đồng thuận: kappa = 82%
- Mỗi ngôn ngữ có 2500 ví dụ trung tính, 750 ví dụ tích cực, và 750 ví dụ tiêu cực
- Đánh giá dựa trên 10-fold cross-validation

Thiết lập thí nghiệm

- Mô hình cascade:
  - Tầng 1: unigram
  - Tầng 2: + discourse + negation
  - Tầng 3: + parsed feature
- SC uni-lang (~ tầng 2) huấn luyện trên toàn bộ dữ liệu
- SC uni-lang-dist bổ sung thêm đặc trưng khoảng cách giữa từ và thực thể
- English: MNB; Ducht: SVM; French: ME

Đánh giá

Vai trò của các đặc trưng

Features	SVM	MNB	ME
Unigrams	85.45%	81.45%	84.80%
Unigrams & BSubjectivity	**86.35%**	83.95%	**87.40%**
Bigrams	85.35%	83.15%	85.40%
Adjectives	75.85%	82.00%	80.30%

Vai trò của bagging

Table 4: Results of the the first layer (English corpus) – 10-fold cross-validation.

Features	Precision neu/not neu	Recall neu/not neu	F-measure neu/not neu
Using bagging	96.05/51.69	62.15/94.06	75.47/66.71
No bagging	88.79/78.07	86.20/81.87	87.48/79.92

Đánh giá (tiếp)

Vai trò của mô hình cascade

(a) English

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	83.30	69.09/85.48/85.93	55.73/82.40/91.84	61.70/83.91/88.79
Cascade with layers 1 and 2	83.10	70.49/87.72/84.61	54.13/79.07/93.00	61.24/83.17/88.61
SC uni-lang	83.03	69.59/86.77/85.08	56.13/79.60/92.12	62.14/83.03/88.46
SC uni-lang-dist	80.23	60.59/78.78/86.57	59.87/82.67/85.60	60.23/80.68/86.08
SC uni	82.73	68.01/85.63/85.53	58.40/78.67/91.24	62.84/82.00/88.29

(b) Dutch

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1,2 and 3	69.03	63.51/53.30/72.20	42.93/31.20/88.20	51.23/39.36/79.40
Cascade with layers 1 and 2	69.80	66.60/58.31/71.66	41.73/29.47/90.32	51.31/39.15/79.92
SC uni-lang	69.05	60.39/52.59/73.63	49.60/33.87/85.44	54.47/41.20/79.10
SC uni-lang-dist	68.85	61.08/54.52/72.20	43.73/30.53/87.88	50.97/39.15/79.27
SC uni	68.18	58.73/49.58/73.24	48.00/31.73/85.16	52.82/38.70/78.75

(c) French

Architecture	Accuracy	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	67.68	50.74/55.88/71.90	27.47/38.67/88.44	35.64/45.71/79.32
Cascade with layers 1 and 2	67.47	52.69/53.96/71.56	26.13/38.13/88.68	34.94/44.69/79.21
SC uni-lang	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni-lang-dist	65.97	47.67/50.33/72.18	30.00/40.67/84.36	36.82/44.99/77.79
SC uni	65.83	45.67/50.82/72.23	28.80/41.33/84.28	35.32/45.59/77.79

Đánh giá (tiếp)

Hiệu năng trên các câu trung tính

(a) English

Architecture	Precision	Recall	F-measure
Layer 1 of the cascade	88.79	86.20	87.48
Layer 1 and 2 of the cascade	84.61	93.00	88.61
Layer 2 of the cascade	85.08	92.12	88.46

(b) Dutch

Architecture	Precision	Recall	F-measure
Layer 1 of the cascade	74.49	82.00	78.07
Layer 1 and 2 of the cascade	71.66	90.32	79.92
Layer 2 of the cascade	73.73	85.88	79.34

(c) French

Architecture	Precision	Recall	F-measure
Layer 1 of the cascade	75.95	81.36	78.56
Layer 1 and 2 of the cascade	71.56	88.68	79.21
Layer 2 of the cascade	72.18	84.36	77.79

Đánh giá (tiếp)

Vai trò của miền lĩnh vực

(a) Car domain

Architecture	Acc	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	70.65	74.04/62.32/71.45	55.78/39.33/87.28	63.62/48.23/78.57
SC uni-lang	70.84	69.67/63.02/72.83	60.22/43.56/84.48	64.60/51.51/78.22
SC uni-lang-dist	70.51	70.23/65.06/71.53	54.00/38.89/87.84	61.06/48.68/78.85
Cascade layers 1, 2 and 3 trained on movie domain	63.95	62.33/48.47/65.72	40.44/17.56/89.12	49.06/25.77/75.65

(b) Movie domain

Architecture	Acc	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	56.88	46.15/31.76/59.85	12.00/12.00/89.2	19.05/17.42/71.64
SC uni-lang	59.77	44.76/48.09/65.29	31.33/33.56/79.44	36.86/39.53/71.67
SC uni-lang-dist	62.05	48.70/51.81/66.04	29.11/31.78/84.80	36.44/39.39/74.26
Cascade layers 1, 2 and 3 trained on car domain	59.40	55.61/36.18/63.98	25.33/23.56/84.56	34.81/28.53/72.85

Đánh giá (tiếp)

Vai trò của đặc trưng cú pháp đối với các câu nhập nhằng

Table 6: Results with regard to the classification of ambiguous positive, negative and neutral sentences (Dutch corpus) that can be parsed – 10-fold cross-validation.

Features	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Unigrams	28.57/100/30.43	6.06/ 7.69/93.33	10.00/14.29/45.90
Unigrams + parse features	33.33/100/32.18	9.09/15.38/93.33	14.29/26.67/47.86

Hiệu năng đối với các câu không chắc chắn

Table 7: Results with regard to the classification of uncertain positive, negative and neutral sentences (English corpus) that can be parsed – 10-fold cross-validation.

Architecture	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	50.48/69.05/57.14	71.62/86.14/11.43	59.22/76.65/19.05
Cascade with layers 1 and 2	53.95/78.48/41.11	55.41/61.39/52.86	54.67/68.89/46.25

Đánh giá (tiếp)

Hiệu năng đối với các câu có nhiều nhiễu

Table 8: Results with regard to the classification of very noisy sentences that diverge from formal language (French corpus) – 10-fold cross-validation.

Architecture	Precision pos/neg/neu	Recall pos/neg/neu	F-measure pos/neg/neu
Cascade with layers 1, 2 and 3	60.87/52.63/83.41	24.28/15.87/97.11	34.71/24.39/89.74
SC uni-lang	61.25/55.00/83.90	28.32/17.46/96.56	38.74/26.51/89.78
SC uni-lang-dist	61.25/55.00/83.90	28.32/17.46/96.56	38.74/26.51/89.78

Phân tích lỗi

Table 9: Error analysis based on examination of 50 misclassified sentences in English, Dutch and French.

Id	Cause	English	Dutch	French	All
1	Features insufficiently known and/or wrong feature connotations	23	21	15	59
2	Ambiguous examples	12	8	8	28
3	Sentiment towards (sub-)entity	3	3	9	15
4	Cases not handled by negation	3	3	4	10
5	Expressions spanning several words	3	5	2	10
6	Understanding of the context or world knowledge is needed	2	2	4	8
7	Domain specific	0	3	3	6
8	Language collocations	2	2	2	6
9	A sentiment feature has multiple meanings	2	1	2	5
10	Language specific	0	2	1	3

Phân tích lỗi (2)

1. Thiếu dữ liệu huấn luyện dẫn đến các câu không mang cảm xúc nhưng bị gán thành tích cực/tiêu cực do có chứa các từ thường xuất hiện trong các câu tích cực/tiêu cực
2. Các câu nhập nhằng
3. Câu mang cảm xúc đối với thực thể khác
4. Câu mang từ phủ định
5. Cảm xúc được thể hiện qua ẩn dụ
6. Cảm xúc phải được suy ra từ ngữ cảnh rộng (văn bản) hoặc tri thức chung

Phân tích lỗi (3)

7. Yêu cầu tri thức về lĩnh vực hẹp
8. Cảm xúc thể hiện qua thành ngữ
9. Cảm xúc thể hiện qua từ đa nghĩa
10. Các vấn đề cụ thể về ngôn ngữ như từ ghép trong Ductch hay dấu trong French

Phân tích lỗi (4)

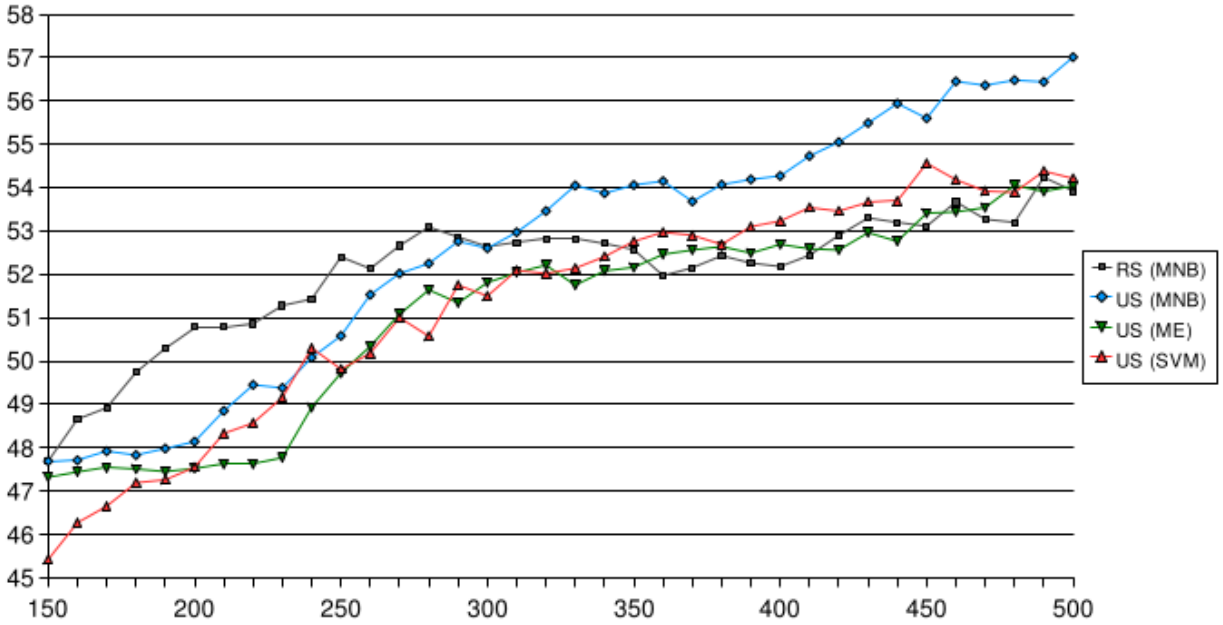
2. *A Good Year is a **fine** example of a **top-notch** director and actor out of their elements in a sappy romantic comedy lacking in …*
3. *certainly more comfortable and rewarding than an Audi Q7 and …*
5. *it is een schot in de roos (this is a shot in the bull's eye) ~ they got it exactly right*
6. *I don t know maybe it s because I was younger back then but Casino Royale felt more like a connect the dots exercise than a Bond movie.*
7. *[...] attention pour avoir une chance de ne pas dormir au bout de 10 minutes, mieux vaut connaître les règles du poker […] (dormir ~ sleep)*
8. *Casino Royale finally hits full-throttle in its second hour but Bond fans will find the movie hit-and-miss **at best***
9. *Not a coincidence–GM used Mercedes' supplier for the new ... the interior plastics and wood trims is REALLY cheap. ... brown seats in a light colored car only make*

A Ferrari is not cheap to buy or run and residual values weaken if you use the car regularly.

Đánh giá học chủ động

- Chọn đến khi đủ 500 ví dụ
- Sử dụng bộ phân loại MNB với tập đặc trưng unigram + discourse + negation
- Đánh giá trên English
- Tập validation: 1703 ví dụ: 1151 trung tính, 274 tích cực, và 72 tiêu cực

Đánh giá học chủ động (2)



	MNB		ME		SVM	
#Ex	RS	US	RS	US	RS	US
150	47.69	47.69	47.32	47.32	45.82	45.82
200	50.79	48.14	49.54	47.52	46.94	47.55
250	52.40	50.58	51.89	49.72	46.83	49.81
300	52.64	52.60	52.38	51.81	47.55	**51.50**
350	52.57	54.06	51.90	52.16	47.86	**52.76**
400	52.18	54.27	52.32	52.69	48.27	**53.23**
450	53.11	**55.60**	52.61	53.41	48.87	**54.56**
500	53.92	**57.01**	52.08	54.04	48.55	**54.21**

Đánh giá học chủ động (3)

Lấy mẫu không chắc chắn vs lấy mẫu ngẫu nhiên

Table 11: Comparison of RS and US for the MNB uncertainty sampling method using seed size 150 and batch size 10. The number after ± is the standard deviation – averaged over 5 runs.

	Accuracy		F-measure pos		F-measure neg	
#Ex	RS	US	RS	US	RS	US
150	68.10±00.39	68.10±00.39	35.05±06.70	35.05±06.70	26.64±03.21	26.64±03.21
200	73.45±01.01	70.23±00.60	36.50±08.75	33.74±08.32	30.97±02.32	27.67±03.06
250	75.88±01.20	74.25±01.36	37.41±09.15	35.02±07.98	33.43±01.40	31.46±03.55
300	77.53±00.88	76.74±01.61	36.96±10.48	37.91±02.95	33.65±02.75	33.20±04.99
350	78.40±01.06	77.79±01.46	38.63±09.60	40.51±03.10	31.30±06.08	34.47±07.12
400	78.46±00.71	78.25±01.59	38.26±10.33	41.06±02.17	30.52±06.44	34.38±06.30
450	79.21±00.98	79.42±01.27	39.30±06.95	42.08±03.98	31.87±05.94	36.62±05.24
500	79.54±00.70	80.06±01.04	40.15±06.19	44.40±03.63	33.30±05.40	38.21±05.97

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict