Tài liệu giảng dạy môn Khai thác dữ liệu Web

KHAI THÁC Ý KIẾN

TS. Nguyễn Ngọc Thảo – ThS. Lê Ngọc Thành

Bộ môn Khoa học Máy tính, FIT HCMUS, VNUHCM

Thành phố Hồ Chí Minh, 02/2019

Nội dung bài giảng

- Bài toán Khai thác ý kiến
- Phân loại ý kiến ở mức văn bản
- Phân loại ý kiến ở mức câu
- Xây dựng tập thuật ngữ ý kiến
- Phân loại ý kiến ở mức khía cạnh
- Khai thác ý kiến so sánh
- Phát hiện ý kiến spam
- Một số vấn đề khác

Bài toán Khai thác ý kiến

Khi nào cần ý kiến người dùng?

★★★★★ **Nice Iphone, 256 GB**

By marco augusto on January 18, 2017

Nice Iphone, 256 GB .Flash drive is excellent to storage more information and to use more applications, the camera is really good, battery life is longer than iphone 6s, the processor is faster than previous models. Is a excellent phone.

★★★★☆ **Good camera,wide screen and apps and games**

By Amazon Customer on November 21, 2016

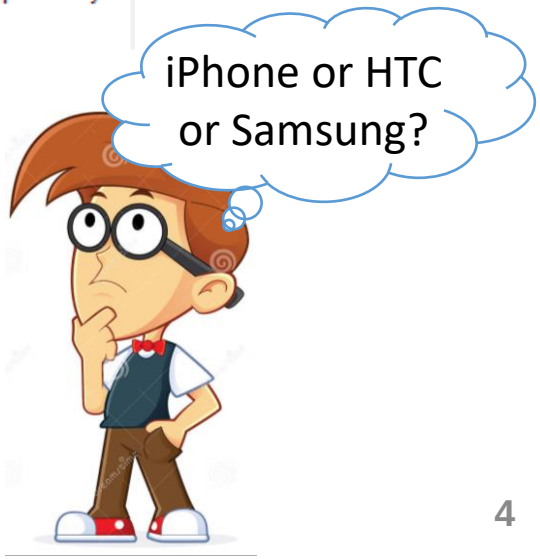
Verified Purchase

The iPhone 7 plus has good camera shoot and amazing 3D Touch and 5.5" screen size but less apps and games. I think the iOS 10.2 will bring more amazing apps and games in the store.

★☆☆☆☆ **Battery run down too fast**

By Amazon Customer on February 9, 2017

The battery runs down very fast. I dont even make calls on it yet nor do i download any social app on it yet it just ran down. Is it that the battery i got with the phone not good?



Khai thác ý kiến là gì?

- **Khai thác ý kiến** (opinion mining, sentiment analysis) phân tích ý kiến, đánh giá, thái độ và cảm xúc của con người về những thực thể, cá nhân, vấn đề, sự kiện, chủ đề và những thuộc tính liên quan của chúng.
- Ý kiến ảnh hưởng chủ yếu đến hành vi của con người → khái niệm quan trọng.
  - Lựa chọn của con người phần lớn dựa theo cách người khác nhìn nhận và đánh giá sự việc.

Ý kiến người dùng

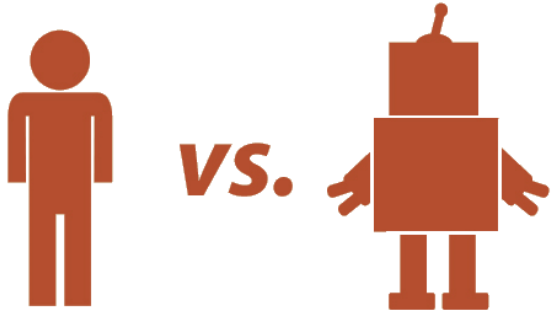
- Nhận xét về sản phẩm tồn tại dồi dào trong các review, forum discussion, blog và mạng xã hội.

Các hình thức tham khảo ý kiến truyền thống như hỏi người thân, tổ chức bình chọn, khảo sát…trở nên không cần thiết.

- Tuy nhiên, giải mã ý kiến người dùng từ một lượng lớn thông tin trong truyền thông xã hội là một thử thách lớn.

Khai thác ý kiến tự động

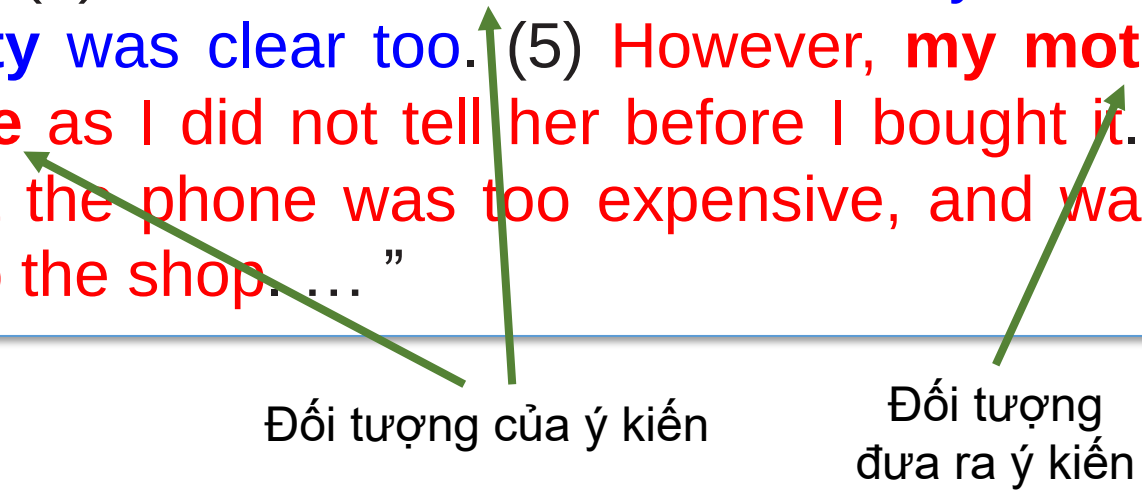
- Khả năng vật lý và nhận thức của con người có giới hạn.
  - Khó nhận diện chính xác trang liên quan và tóm tắt ý kiến trong đó.
  - Không thể giữ tính nhất quán khi xử lý **thủ công** lượng thông tin lớn.
  - Việc phân tích và đánh giá mang tính **chủ quan**, thường quan tâm nhiều đến ý kiến cùng sở thích cá nhân.



- Hệ thống tóm tắt và khai thác ý kiến **tự động** có thể thực thi tác vụ một cách **khách quan**.

Ví dụ minh họa

"(1) I bought an iPhone a few days ago. (2) It was such a nice **phone**. (3) The **touch screen** was really cool. (4) The **voice quality** was clear too. (5) However, **my mother** was mad with **me** as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop. … "

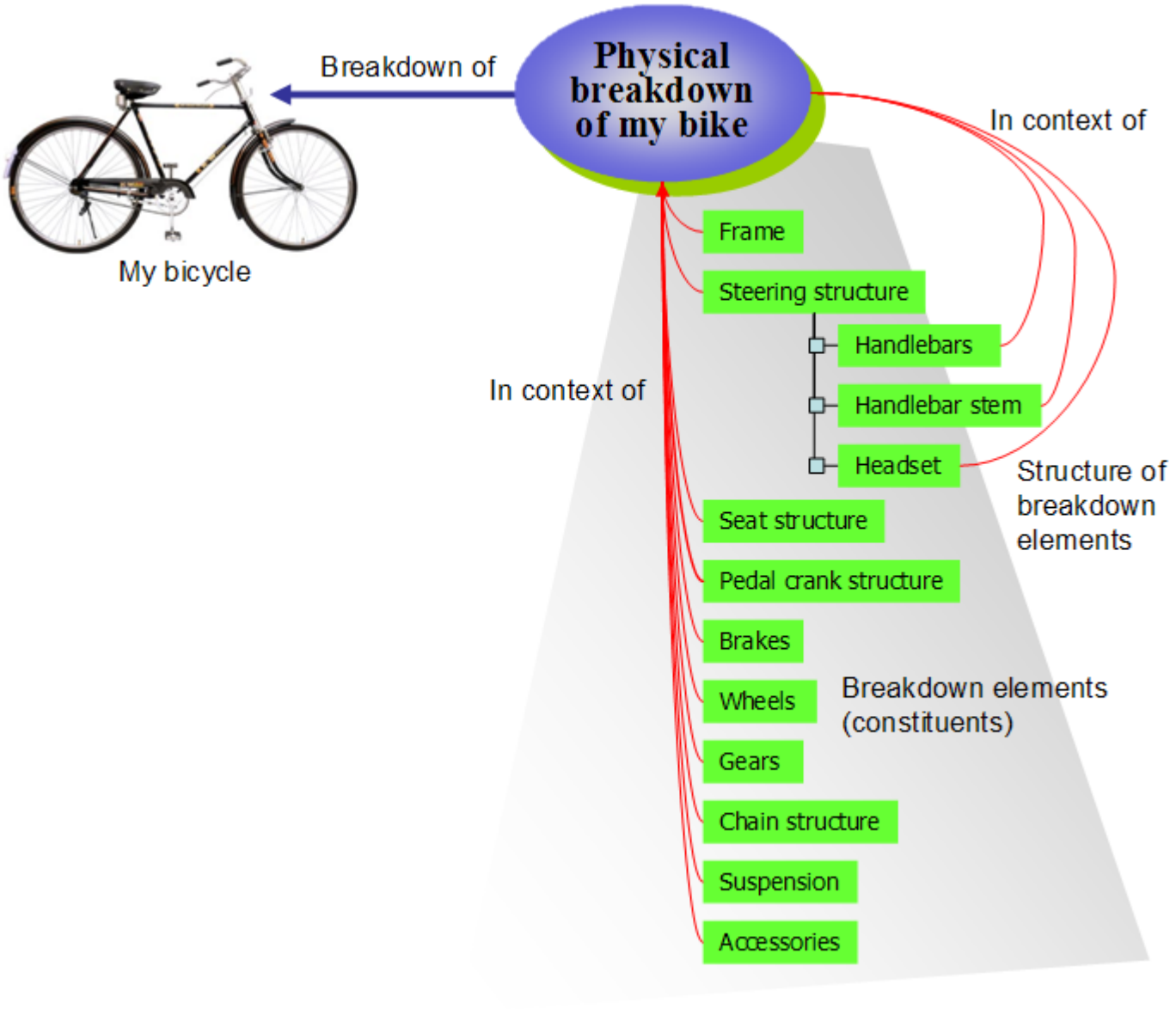


- Chúng ta cần khai thác gì từ bình luận trên?
  - Ý kiến/cảm xúc tích cực/tiêu cực.
  - Đối tượng đưa ra ý kiến hoặc được đề cập trong ý kiến.

Khái niệm cơ bản: Thực thể

- **Thực thể** (entity) chỉ đối tượng mục tiêu được đánh giá.
  - Sản phẩm, dịch vụ, con người, sự kiện hay chủ đề.
- Thực thể là một hệ thống phân cấp, ký hiệu $\boldsymbol{e}$: $(\boldsymbol{T}, \boldsymbol{W})$.
  - T là cấu trúc phân cấp biểu diễn các bộ phận trong $\boldsymbol{e}$ và W là tập thuộc tính (attribute) của $\boldsymbol{e}$.
  - Nút gốc chứa tên của thực thể và nút trong chỉ bộ phận (component hay sub-component) của thực thể, mỗi nút có tập thuộc tính riêng.
  - Liên kết giữa các nút là quan hệ *part-of*.
- Ý kiến có thể được phát biểu trên bất kỳ nút nào và bất kỳ thuộc tính nào của nút.

Ví dụ về thực thể



Ví dụ về thực thể

I do not like iPhone.

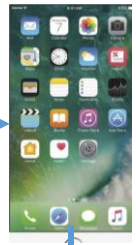
voice quality,
size, weight,…

The voice quality of iPhone is clear.

part-of

battery life,
battery size,…

Battery run down too fast.



resolution,
brightness,…

…

Khái niệm cơ bản: Khía cạnh

- **Khía cạnh** (aspect) của thực thể e bao hàm các bộ phận và thuộc tính của e.
  - Tên khía cạnh (aspect name) là tên đặt cho khía cạnh của thực thể.
  - Biểu diễn khía cạnh (aspect expression) là từ/cụm từ thật sự xuất hiện trong văn bản để chỉ khía cạnh.

    Name: voice quality

    Expression: sound, voice, voice quality

- Tương tự, thực thể cũng có tên thực thể (entity name) và các biểu diễn thực thể (entity expression).
  - Ví dụ, Motorola có thể được biểu diễn là “Moto,” “Mot,” và “Motorola”.

Khái niệm cơ bản: Khía cạnh

- Aspect expression thường là *noun* hay *pronoun*, cũng có thể là *verb*, *verb phrase*, *adjective* hay *adverb*.
- Explicit aspect expression: noun hay noun phrase
  - Ví dụ, “sound” trong “The sound of this phone is clear.”
- Implicit aspect expression: các thể còn lại
  - Ví dụ, “large” trong “This phone is too large.” chỉ kích cỡ của điện thoại, expensive (giá cả), reliably (độ tin cậy).
  - “fit in pockets” trong “This phone will not easily fit in pockets.”, ám chỉ kích cỡ và/hoặc hình dạng.

Thực thể và khía cạnh

- Thực thể thường được đơn giản hóa trong thực tế.
  - Nghiên cứu văn bản ở bất kỳ mức độ chi tiết nào là rất phức tạp.
  - Người dùng thông thường không thể sử dụng biểu diễn phân cấp.
- Cây đơn giản hóa chỉ gồm hai cấp: nút gốc – thực thể và nút con – các khía cạnh khác nhau của thực thể.

Khái niệm cơ bản: Người cho ý kiến

- **Người cho ý kiến** (opinion holder, opinion source) là cá nhân hay tổ chức đưa ra ý kiến.
  - Đối với các trang/blog bình luận sản phẩm, người cho ý kiến thường là chủ của bài đăng.
  - Đối tượng đưa ra ý kiến trong các bài báo tin tức thường có danh tính rõ ràng và do đó có độ quan trọng cao.

Khái niệm cơ bản: Phân loại ý kiến

- Có hai loại ý kiến chính: **ý kiến thông thường** (regular opinion) và **ý kiến so sánh** (comparative opinion).
- Ý kiến thông thường: là đánh giá (tích cực hay tiêu cực) từ người cho ý kiến, về thực thể hay khía cạnh của thực thể.
  - Ví dụ, Iphone X is very expensive.
- Ý kiến so sánh: diễn đạt mối quan hệ giống/khác nhau giữa hai hay nhiều thực thể dựa trên những khía cạnh chung, và thường đi kèm sự ưu tiên của người cho ý kiến.
  - Ví dụ, Iphone X is more expensive than Samsung phone.

Khái niệm cơ bản: khuynh hướng ý kiến

- **Khuynh hướng ý kiến** (opinion orientations): tích cực (positive), tiêu cực (negative), hoặc trung lập (neutral).
- Tên gọi khác: sentiment orientation, semantic orientation, hoặc polarity.

Phân biệt ý kiến và cảm xúc

- Cảm xúc (emotion) là các cảm giác và suy nghĩ chủ quan.
  - Cảm xúc cơ bản: yêu, vui, ngạc nhiên, giận, buồn bã và sợ hãi.
  - Mỗi cảm xúc có các cường độ khác nhau và có thể chia thứ cấp.

- Độ mạnh của ý kiến tùy thuộc vào cường độ của cảm xúc.
- Tuy nhiên, cảm xúc và ý kiến không tương đương nhau.
  - Câu chứa ý kiến chưa chắc hàm chứa cảm xúc và ngược lại.
  - Ví dụ, “The voice of this TV is clear.” và “I’m so surprised to see you.”

Mô hình hóa ý kiến

- Ý kiến là một quintuple $(\boldsymbol{e_i, a_{ij}, oo_{ijkl}, h_k, t_l})$, trong đó
  - e_i: tên của thực thể.
  - a_{ij}: khía cạnh của e_i (GENERAL nếu xét toàn bộ thực thể).
  - oo_{ijkl}: khuynh hướng ý kiến về khía cạnh a_{ij}, được thể hiện bằng {positive, negative, neutral}, hoặc các mức cường độ khác nhau.
  - h_k: người cho ý kiến.
  - t_l: thời điểm mà h_k đưa ra ý kiến.
- Năm thành phần trong quintuple phải tương ứng với nhau để tránh gán ý kiến vào sai thực thể hay sai khía cạnh.
  - Ý kiến oo_{ijkl} được đưa ra bởi người cho ý kiến h_k về khía cạnh a_{ij} của thực thể e_i tại thời điểm t_l.

Nhận xét về mô hình ý kiến

- Các thành phần trong quintuple là yếu tố cơ bản nhất.
  - Ví dụ, ý kiến “The picture quality is great.” ít có giá trị sử dụng vì không biết “picture quality” của thực thể nào.
- Tuy nhiên, không phải lúc nào cũng cần đủ năm thành phần.
  - Ví dụ, không cần quan tâm h_k khi tổng hợp ý kiến từ nhiều người.
- Có thể thêm thành phần khác vào mô hình, tùy ngữ cảnh.
  - Ví dụ, thêm giới tính (nam, nữ) của người cho ý kiến trong ngữ cảnh tiếp thị sản phẩm.

Nhận xét về mô hình ý kiến

- Nền tảng chuyển đổi văn bản phi cấu trúc thành dữ liệu cấu trúc → phân tích ý kiến một cách định tính và định lượng.
  - Phân tích bằng DBMS và công cụ OLAP (slice/dice)
- Ý kiến có thể được chia thành **ý kiến trực tiếp** (direct opinion) và **ý kiến gián tiếp** (indirect opinion).
  - "The voice quality of this phone is great." → ý kiến trực tiếp.
  - "After taking this drug, my hand felt much better." → ý kiến gián tiếp thể hiện nhận xét tích cực về "drug" đối với "my hand".

Khai thác ý kiến theo khía cạnh

- Mô hình thực thể
  - Thực thể $\boldsymbol{e_i}$ chứa tập hữu hạn các khía cạnh $\boldsymbol{A_i} = \{\boldsymbol{a_{i1}}, \boldsymbol{a_{i2}}, \dots, \boldsymbol{a_{in}}\}$.
  - Mỗi $\boldsymbol{a_{ij}} \in \boldsymbol{A_i}$ được thể hiện bằng phần tử trong tập biểu diễn khía cạnh $\boldsymbol{AE_{ij}} = \{\boldsymbol{ae_{i1}}, \boldsymbol{ae_{i2}}, \dots, \boldsymbol{ae_{im}}\}$
  - Bản thân $\boldsymbol{e_i}$ cũng được biểu thị bằng phần tử trong tập biểu diễn thực thể $\boldsymbol{OE_i} = \{\boldsymbol{oe_{i1}}, \boldsymbol{oe_{i2}}, \dots, \boldsymbol{oe_{is}}\}$.
- Mô hình tài liệu giữ ý kiến
  - Tài liệu giữ ý kiến $\boldsymbol{d}$ chứa ý kiến về tập thực thể $\{e_1, e_2, \dots, e_r\}$ từ các người cho ý kiến $\{h_1, h_2, \dots, h_p\}$.
  - Ý kiến về thực thể e_i được phát biểu trên chính e_i và một tập con A_{id} từ các khía cạnh của e_i.

Khai thác ý kiến theo khía cạnh

- Cho một tập hợp các tài liệu giữ ý kiến D, tìm mọi quintuple $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ trong D.

- Các tác vụ đã nêu đều chưa được giải quyết triệt để.
- Một số thành phần thông tin không được đề cập tường minh trong câu do cách sử dụng đại từ, quy ước ngôn ngữ và ngữ cảnh.
- Việc đảm bảo năm thành phần thông tin trong một ý kiến tương ứng với nhau cũng là thử thách lớn.

Ví dụ về khai thác ý kiến

Posted by: bigXyz on Nov-4-2010: (1) I bought a Motorola phone and my girlfriend bought a Nokia phone yesterday. (2) We called each other when we got home. (3) The voice of my Moto phone was unclear, but the camera was good. (4) My girlfriend was quite happy with her phone, and its sound quality. (5) I want a phone with good voice quality. (6) So I probably will not keep it.

- (Motorola, voice_quality, negative, bigXyz, Nov-4-2010)
- (Motorola, camera, positive, bigXyz, Nov-4-2010)
- (Nokia, GENERAL, positive, bigXyz's girlfriend, Nov-4-2010)
- (Nokia, voice_quality, positive, bigXyz's girlfriend, Nov-4-2010)

Tóm tắt ý kiến theo khía cạnh

Tóm tắt ý kiến theo khía cạnh

- Aspect-based opinion summary
- Đây là bài toán cần thiết để tổng hợp ý kiến từ một lượng lớn người cho ý kiến bằng cách sử dụng quintuple.

Cellular phone 1:

Aspect: **GENERAL**

Positive:	125	<individual review sentences>
Negative:	7	<individual review sentences>

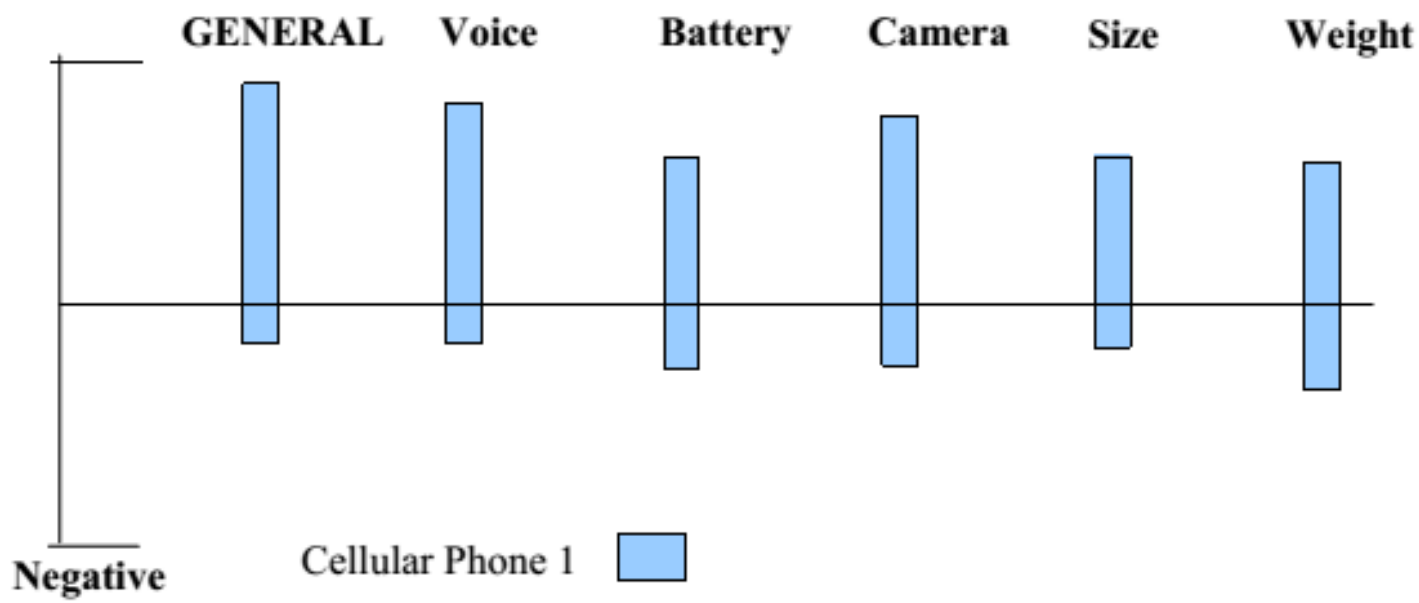
Aspect: **Voice quality**

Positive:	120	<individual review sentences>
Negative:	8	<individual review sentences>

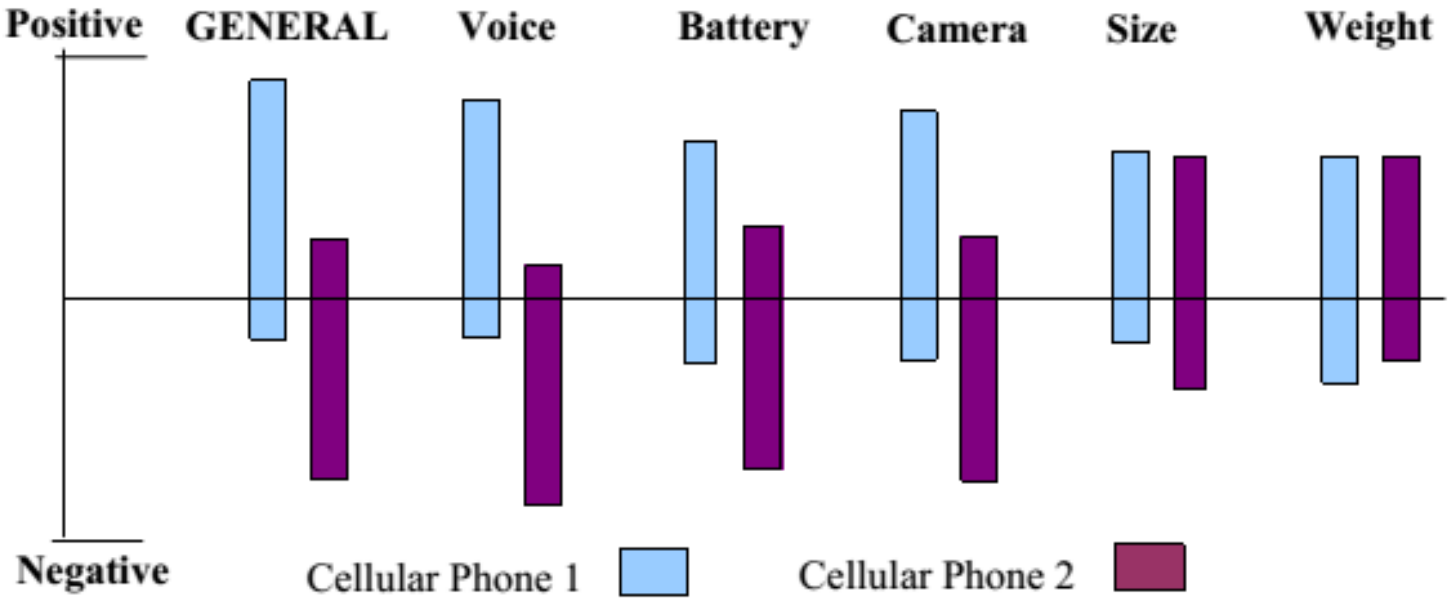
Aspect: **Battery**

Positive:	80	<individual review sentences>
Negative:	12	<individual review sentences>

...



(A) Visualization of aspect-based summary of opinions on a cellular phone



(B) Visual opinion comparison of two cellular phones

Tóm tắt ý kiến theokhía cạnh

- Hướng tiếp cận truyền thống: phát sinh tóm tắt văn bản trực tiếp từ những ý kiến.
  - Chỉ định tính chứ không định lượng, không thích hợp cho phân tích.
  - Ví dụ: tóm tắt truyền thống tạo ra câu “Most people do not like this product.”, phân tích định lượng cho thấy 60% không thích sản phẩm và 40% thích sản phẩm.
- Hướng tiếp cận khác: phát sinh tóm tắt văn bản từ kết quả khai thác dữ liệu.
  - Sử dung template để phát sinh câu ngôn ngữ tự nhiên từ đồ thị.
  - Ví dụ: “Most people are positive about cellular phone 1 and negative about cellular phone 2.”

Phân loại ý kiến mức văn bản

Định nghĩa bài toán

- Document-level sentiment classification, phân loại một tài liệu giữ ý kiến là phát biểu tích cực hay tiêu cực.
- **Phát biểu bài toán**: Cho tài liệu giữ ý kiến $\boldsymbol{d}$ đánh giá về thực thể $\boldsymbol{e}$, xác định khuynh hướng ý kiến $\boldsymbol{oo}$ trên $\boldsymbol{e}$ trong quintuple $(e, GENERAL, oo, h, t)$.
  - Giả sử e, h, và t đã biết hoặc không liên quan.
- **Ràng buộc**: Tài liệu d phát biểu các ý kiến về thực thể đơn e và những ý kiến này chỉ từ một người cho ý kiến h.

Phân loại bằng học có giám sát

- Phân loại ý kiến được định nghĩa như một bài toán học có giám sát với ba lớp – positive, negative và neutral.
- Dữ liệu huấn luyện và kiểm thử lấy từ bình luận sản phẩm.
  - Sử dụng điểm số xếp hạng của tác giả (ví dụ 1 – 5 sao) để tạo nhãn: tích cực (4 – 5 sao), tiêu cực (1 – 2 sao), và trung lập (3 sao)

- Có thể sử dụng bất kỳ phương pháp học có giám sát nào.
  - Ví dụ: naïve Bayesian, SVM, v.v.

Học có giám sát: Lựa chọn đặc trưng

- Từ và tần số từ
  - Từ đơn hoặc từ n-gram.
  - Có thể áp dụng TF-IDF hay xét thứ tự từ.

Word pair	TF Rank	TF-IDF Rank
fresh fish	1	2
great sushi	23	19
delicious sushi	(n/a)	40
good food	24	(n/a)
good place	11	(n/a)
delicious sashimi	(n/a)	21
good quality	38	(n/a)
good sushi	7	9
sushi place	6	14
sashimi platter	(n/a)	22
delicious fish	(n/a)	39
good service	29	(n/a)
fresh quality	(n/a)	33
grilled salmon	(n/a)	36
Japanese restaurant	18	(n/a)

Novati, M., Trusty, A., Truong, K.N., & Yatani, K. (2011). *Analysis of Adjective-Noun Word Pair Extraction Methods for Online Review Summarization*. IJCAI.

Học có giám sát: Lựa chọn đặc trưng

- Part of speech
  - Tính từ là chỉ thị quan trọng về ý kiến → xử lý đặc biệt.
- Từ và cụm từ ý kiến
  - Từ ý kiến dùng để tạo phát biểu tích cực (ví dụ beautiful, wonderful, good và amazing) hay tiêu cực (ví dụ bad, poor và terrible).
  - Đa số là tính từ và trạng từ, một số là danh từ (ví dụ rubbish, junk và crap) và động từ (ví dụ hate và like).
  - Cụm từ ý kiến hay thành ngữ: “cost someone an arm and a leg”.
- Sự phủ định
  - Thường thay đổi khuynh hướng ý kiến, ví dụ, “I don’t like this camera”.
  - Không phải cứ xuất hiện từ phủ định là có nghĩa phủ định, ví dụ, “not” trong “not only…but also”.

Học có giám sát: Lựa chọn đặc trưng

- Luật biểu diễn ý kiến
  - Ý kiến có thể được biểu diễn theo nhiều cách khác nhau.

1. POSITIVE ::= P
2. | PO
3. | orientation shifter N
4. | orientation shifter NE
5. NEGATIVE ::= N
6. | NE
7. | orientation shifter P
8. | orientation shifter PO

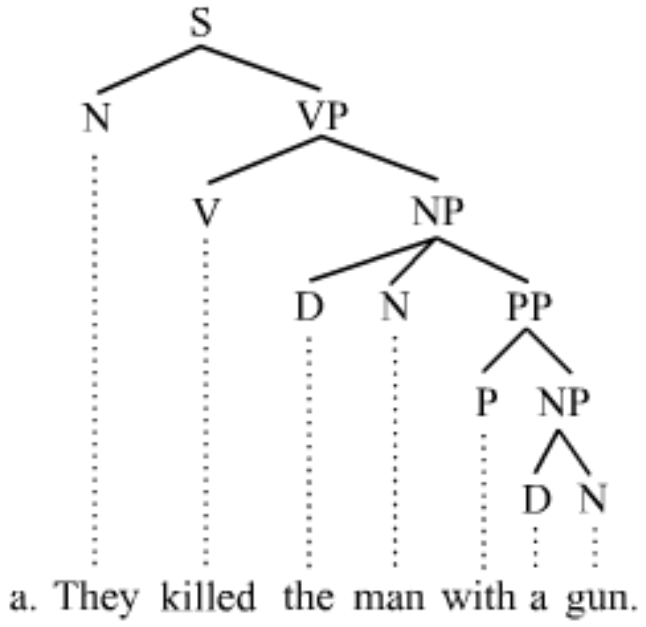
  - Một số cách biểu diễn không cần từ hay cụm từ ý kiến.

"This drug causes low (or high) blood pressure"

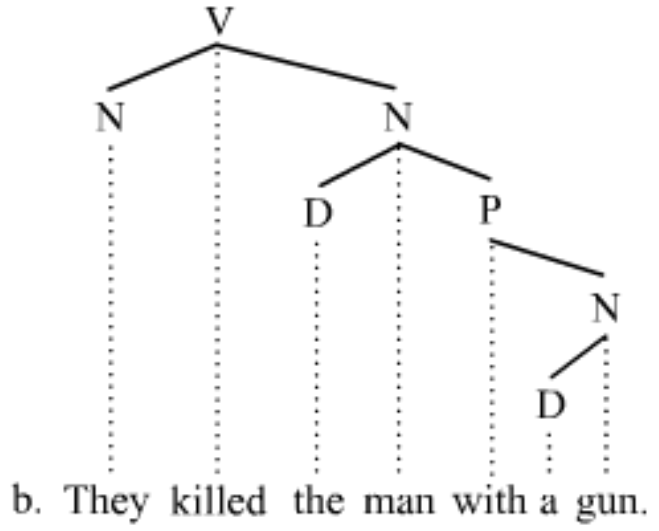
"This drug causes my blood pressure to reach 200."

Học có giám sát: Lựa chọn đặc trưng

- Phụ thuộc cú pháp
  - Các đặc trưng dựa trên phụ thuộc từ được phát sinh bằng cách phân tích cú pháp hay cây phụ thuộc từ.



Phrase structure grammar



Dependency grammar

Các bài toán phân tích ý kiến khác

- Dự đoán điểm đánh giá (ví dụ 1 – 5 sao) cho bài bình luận
  - Bài toán hồi quy, điểm đánh giá có thứ tự
- Thích nghi ngữ cảnh (transfer learning, domain adaptation)
  - Bài toán phân loại ý kiến nhạy cảm với ngữ cảnh mà dữ liệu huấn luyện được rút trích.
  - Những lĩnh vực khác nhau có cách xây dựng câu khác biệt.
  - Cùng một từ nhưng trong lĩnh vực này mang nghĩa tích cực nhưng trong lĩnh vực khác có thể mang nghĩa tiêu cực.

Học không giám sát: Công trình tiêu biểu

- Turney 2002. phân loại dựa trên một số cụm từ cú pháp cố định có khả năng cao được sử dụng để diễn đạt ý kiến.
  - Turney, P. *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002), 2002.
- Thuật toán sử dụng kỹ thuật **part-of-speech (POS) tagging**.
- Phân loại bình luận là *recommended* và *not recommended.*
- Độ chính xác đạt 84% trên các bình luận về xe hơi và 64% trên bình luận phim ảnh.

Table 11.1. Penn Treebank Part-Of-Speech (POS) tags

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Học không giám sát: Công trình tiêu biểu

- Bước 1: rút trích cụm từ chứa tính từ hoặc trạng từ.

	First word	Second word	Third word (not extracted)
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

  - Sử dụng từ đơn không đủ ngữ cảnh để biết khuynh hướng ý kiến → rút trích hai từ liền nhau, một tính từ/trạng từ và một từ ngữ cảnh.
  - Ví dụ, “unpredictable” trong “unpredictable steering” (negative) và trong “unpredictable plot” (positive).

- Bước 2: Ước lượng khuynh hướng ngữ nghĩa của cụm từ được tính theo pointwise mutual information (PMI).

$$PMI(term_1, term_2) = \log_2\left(\frac{\Pr(term_1 \wedge term_2)}{\Pr(term_1)\Pr(term_2)}\right)$$

  - $\Pr(term_1 \wedge term_2)$: xác suất đồng hiện của $term_1$ và $term_2$

Học không giám sát: Công trình tiêu biểu

- Khuynh hướng ý kiến của một cụm từ được xác định dựa trên sự liên kết của nó với từ tham chiếu tích cực "**excellent**" và từ tiêu cực "**poor**".

$$SO(phrase) = PMI(phrase, \text{excellent}) - PMI(phrase, \text{poor})$$

- Các giá trị xác suất được tính bằng cách truy vấn trên cỗ máy tìm kiếm và ghi nhận số tài liệu liên quan (hits).
  - Turney sử dụng AltaVista, có hỗ trợ toán tử NEAR, giới hạn tìm kiếm trong tài liệu có chứa hai từ nằm cách nhau trong vòng 10 từ theo cả hai hướng.
- Công thức SO được viết lại như sau

$$SO(phrase) = \log_2 \left(\frac{hits(\text{phrase } NEAR \text{ excellent}) hits(\text{poor})}{hits(\text{phrase } NEAR \text{ poor}) hits(\text{excellent})} \right)$$

  - $hits(query)$ là số $hits$ trả về, cộng 0.01 vào $hits$ để tránh "division by zero".

- Bước 3: Cho trước một bình luận, tính SO trung bình của mọi cụm từ trong bình luận và phân loại – *recommended* ($SO > 0$) và *not recommended* ($SO < 0$)

Phân loại ý kiến mức văn bản

- **Ưu điểm**
  - Tự động xác định khuynh hướng ý kiến phổ biến về một thực thể, chủ đề hay sự kiện.
- **Khuyết điểm**
  - Không cung cấp chi tiết về khía cạnh nào của thực thể được thích hay không thích
  - Không dễ áp dụng cho những bài viết không phải dạng bình luận (trong diễn đàn và các bài viết blog), vì chúng thường đánh giá nhiều thực thể cùng một lúc

I am very happy with the product. 

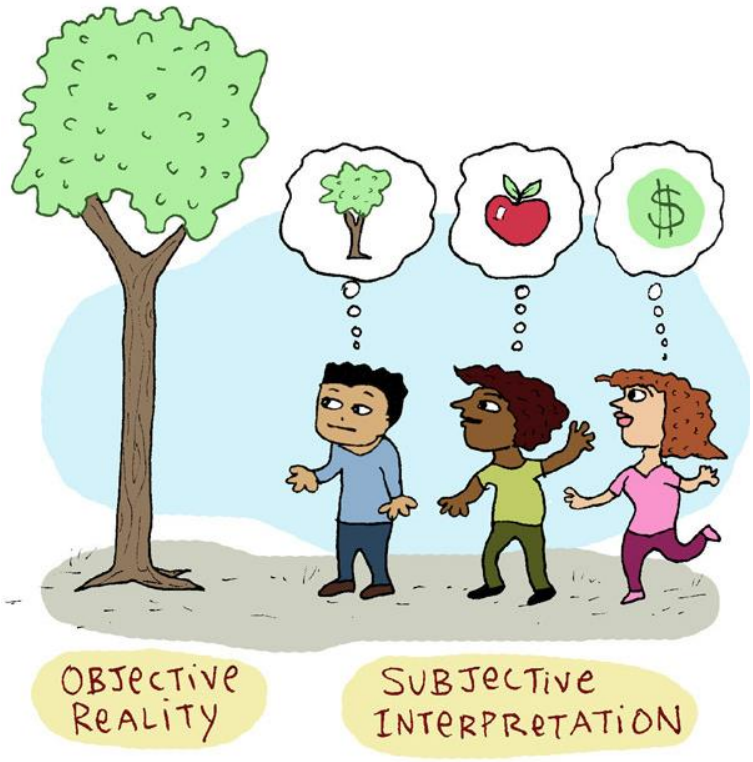
The delivery was on time. 

However, the installation is a little bit tricky. 

Phân loại ý kiến mức câu

Khái niệm chủ quan (subjectivity)

- **Câu khách quan** (objective sentence) trình bày thông tin có thật về thế giới.
- **Câu chủ quan** (subjective sentence) thể hiện cảm giác, góc nhìn hay niềm tin của cá nhân.
- Bài toán xác định câu là chủ quan hay khách quan được gọi là **subjectivity classification**.

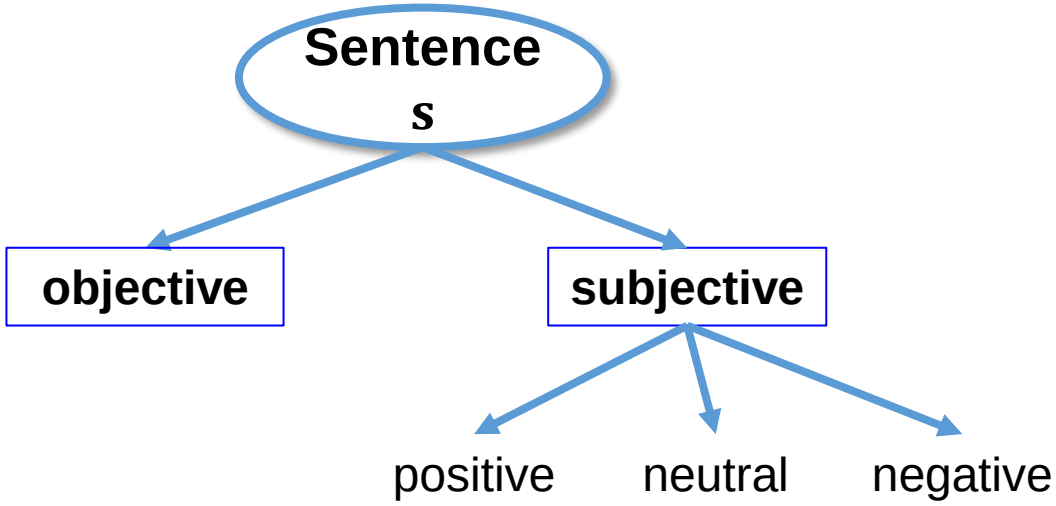


Khái niệm chủ quan (subjectivity)

- Biểu diễn chủ quan có nhiều hình thức.
  - Ý kiến, luận điệu, mong muốn, niềm tin, nghi ngờ và suy đoán.
- Câu chủ quan chưa chắc chứa ý kiến.
  - Ví dụ, “I want a phone with good voice quality.” là chủ quan nhưng không chứa ý kiến tích cực/tiêu cực về bất kỳ điều gì.
- Câu khách quan có thể ngầm chỉ ý kiến.
  - Ví dụ, “The earphone broke in two days.” ám chỉ ý kiến tiêu cực.

Phân loại ý kiến ở mức câu

- Cho trước câu s. Ta cần thực hiện hai tác vụ con như sau.
- Phân loại tính chủ quan
  - Xác định câu s là câu chủ quan hay câu khách quan.
- Phân loại ý kiến ở mức câu
  - Nếu s là câu chủ quan, xác định s thể hiện ý kiến tích cực, tiêu cực hay trung lập.



Phân loại ý kiến ở mức câu

- Chỉ phân loại câu chứa ý kiến tích cực hay tiêu cực, không cho biết thực thể hay khía cạnh nào là mục tiêu của ý kiến
  - Quintuple (e, a, oo, h, t) không được dùng trong định nghĩa.
- Hai tác vụ con trong bài toán đóng vai trò quan trọng.
  (1) Loại bỏ những câu không có ý kiến.
  (2) Sau khi đã biết thực thể và khía cạnh nào được đề cập trong câu, xác định ý kiến về đối tượng tương ứng là tích cực hay tiêu cực.

Phân loại ý kiến ở mức câu

- **Ràng buộc:** Câu diễn đạt một ý kiến đơn lẻ từ một người cho ý kiến.
- Chỉ thích hợp với câu đơn giản có một ý kiến, trong khi câu phức có thể chứa nhiều hơn một ý kiến.
  - Ví dụ, "The picture quality of this camera is amazing and so is the battery life, but the viewfinder is too small for such a great camera."
    - Câu chứa ý kiến tích cực đối với "picture quality", "battery life" và toàn bộ thực thể biểu diễn bởi "this camera" (tức là GENERAL).
    - Câu chứa ý kiến tiêu cực đối với "viewfinder".

Một số công trình tiêu biểu

- Có thể áp dụng phương pháp học có giám sát truyền thống.
  - Ví dụ, Wiebe, J. 1999. phân loại tính chủ quan bằng naïve Bayesian.
- Gán nhãn cho một số lượng lớn mẫu huấn luyện.
  - Riloff, E. and Wiebe, J. 2003 sử dụng hai bộ phân lớp có độ chính xác cao, HP-Subj và HP-Obj, để tự động xác định một số câu chủ quan và khách quan ban đầu.
  - Câu vừa rút trích được dùng để học mẫu theo khuôn cú pháp như:

Syntactic template	Example pattern
<subj> passive-verb	<subj> was satisfied
<subj> active-verb	<subj> complained
active-verb <dobj>	endorsed <dobj>
noun aux <dobj>	fact is <dobj>
passive-verb prep <np>	was worried about <np>

  - Mẫu học được sẽ giúp xác định thêm câu khách quan và chủ quan.

Một số công trình tiêu biểu

- Yu, H. and Hatzivassiloglou, V. 2003. phân loại tính chủ quan bằng học có giám sát và phân loại ý kiến bằng học không giám sát.
  - Tương tự như trong phân loại mức văn bản nhưng với nhiều từ hạt giống hơn và dùng hàm điểm số log-likehood ratio.
- Hatzivassiloglou, V. and Wiebe, J. 2000. xét tính từ chia thành cấp bậc.
- Gamon, M., A. Aue, S. Corston-Oliver, and E. Ringger. 2005. sử dụng học bán giám sát.
- Ngoài ra, còn có các mô hình nhận diện một số loại ý kiến đặc biệt.
- Mệnh đề trong câu cũng có thể chứa các ý kiến khác nhau.
  - Wilson, T., J. Wiebe, and R. Hwa. 2004. phân loại mệnh đề trong câu theo *độ mạnh* của ý kiến trong mệnh đề (neutral, low, medium, high).
  - Wilson, T., J. Wiebe, and P. Hoffmann. 2005. xét từ tác động ngữ cảnh ý kiến như phủ định (ví dụ not và never) và đối lập (ví dụ but và however).

Khai thác ý kiến theo câu

- Câu chủ quan chứa ý kiến chỉ chiếm một phần trong tập các câu ý kiến, trong khi câu khách quan cũng hàm chứa ý kiến.
- Do đó, cần khai thác ý kiến từ cả những câu chủ quan và khách quan.

- Ý kiến còn có thể nằm trong mức sâu hơn, đó là cụm từ.
  - Ví dụ, “Lightweight is the only good point of this disappointing phone.” $\rightarrow$ tích cực với “weight” nhưng tiêu cực với “phone”.

Positive Sentiment - Word Cloud

Negative Sentiment - Word Cloud

Xây dựng tập thuật ngữ ý kiến

Thuật ngữ ý kiến

- **Từ ý kiến tích cực** thể hiện những trạng thái mong muốn.
  - Ví dụ, beautiful, wonderful, good và amazing.
- **Từ ý kiến tiêu cực** thể hiện trang thái không mong muốn.
  - Ví dụ, bad, poor và terrible.
- **Cụm từ ý kiến và thành ngữ**
  - Ví dụ, cost someone an arm and a leg.
- Các từ khóa: opinion lexicon, polar words, opinion-bearing words, hay sentiment words.

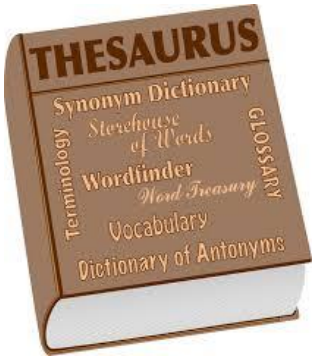
Phân loại thuật ngữ ý kiến

- **Thể cơ sở** (base type)
  - Các từ ví dụ trong slide trước.
- **Thể so sánh** (comparative type) diễn đạt ý kiến so sánh tương quan hoặc tuyệt đối.
  - Ví dụ, better, worse, best, worst là các thể so sánh của tính từ gốc good và bad.
  - Thể so sánh không phát biểu ý kiến trực tiếp về một thực thể mà so sánh nhiều thực thể với nhau.
    - Ví dụ, “Car-x is better than Car-y.” không chỉ Car-x và Car-y tốt hay xấu.

Xây dựng tập thuật ngữ ý kiến

- Hướng tiếp cận thủ công (manual approach)
  - Tốn thời gian, chỉ tham gia vào bước kiểm tra cuối cùng để sửa lỗi do các phương pháp tự động gây ra.
- Hướng tiếp cận tự động (automated approach)
  - Tiếp cận dựa trên từ điển (dictionary-based approach).
  - Tiếp cận dựa trên ngữ liệu (corpus-based approach).

VS.



VS.

Hướng tiếp cận dựa trên từ điển

- Bootstrapping, sử dụng tập từ ý kiến hạt giống có kích thước nhỏ và một từ điển trực tuyến (ví dụ WordNet).
- Tập hợp thủ công một số từ ý kiến đã biết khuynh hướng và phát triển tập này nhiều lần bằng cách bổ sung từ đồng/trái nghĩa lấy từ WordNet.
- Kiểm tra thủ công sau khi thủ tục kết thúc để hiệu chỉnh lỗi.

- Không thể tìm từ ý kiến có khuynh hướng đặc trưng theo ngữ cảnh và lĩnh vực.
  - Ví dụ, “quiet” đối với speaker phone là tiêu cực, nhưng “quiet” đối với xe hơi là tích cực.

Hướng tiếp cận dựa trên ngữ liệu

- Căn cứ vào mẫu cú pháp hoặc mẫu đồng hiện, và tập từ hạt giống ý kiến để tìm các từ khác trong kho ngữ liệu lớn.
- Sentiment consistency: nhận diện tính từ ý kiến (và khuynh hướng) khác dựa vào ràng buộc ngôn ngữ trên từ nối.
  - AND, OR, BUT, EITHER–OR, và NEITHER–NOR
  - Các tính từ liên hiệp bởi AND thường có cùng khuynh hướng, vi dụ, “This car is beautiful and spacious.”

Xây dựng dựa trên ngữ liệu

- Có thể tìm từ ý kiến trong lĩnh vực và ngữ cảnh cụ thể.
- Cùng một từ trong cùng lĩnh vực cũng có thể có khuynh hướng khác nhau trong từng ngữ cảnh.
  - Ví dụ, từ “long” trong “The battery life is long” (positive) và trong “The time taken to focus is long” (negative).
- Không hiệu quả như hướng tiếp cận dựa trên từ điển vì khó chuẩn bị tập ngữ liệu lớn bao phủ mọi từ trong ngôn ngữ.

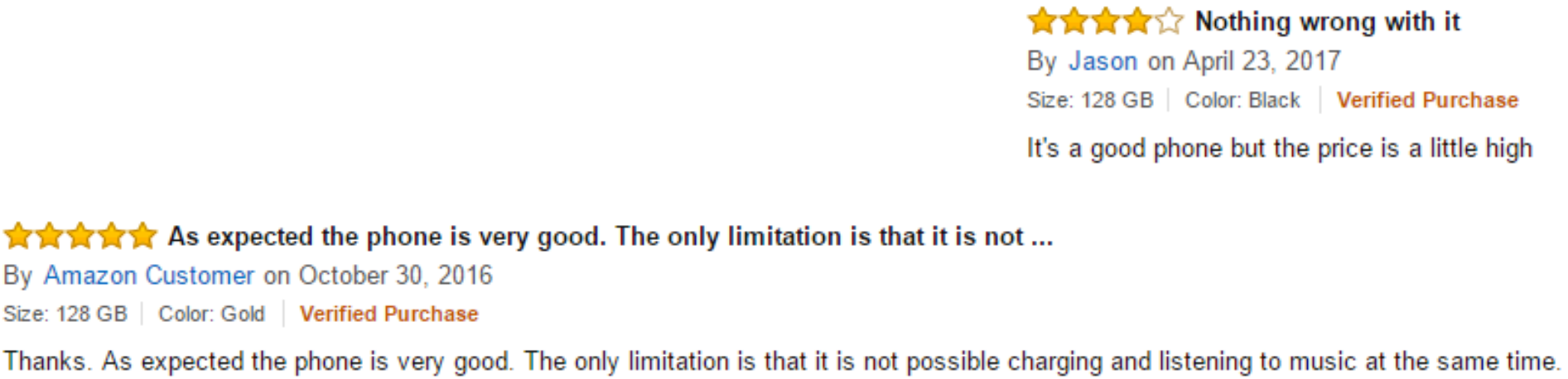
Nhận xét về tập thuật ngữ ý kiến

- Đưa ra một thuật ngữ ý kiến khác với việc xác định khuynh hướng ý kiến của từ/cụm từ trong câu cụ thể.
- Từ xuất hiện trong tập thuật ngữ ý kiến không nhất thiết phải biểu thị ý kiến trong câu.
  - Ví dụ, “I am looking for a good health insurance.” → “good” không thể hiện ý kiến tích cực hay tiêu cực về loại bảo hiểm nào.
- Từ ý kiến không phải là biểu diễn duy nhất có chứa ý kiến.

Phân loại ý kiến theo khía cạnh

Phân loại ý kiến ở mức khía cạnh

- Một tài liệu cho ý kiến tích cực về thực thể không có nghĩa là tác giả ủng hộ mọi khía cạnh của thực thể đó, và ngược lại.



- Phân loại ý kiến ở mức văn bản và mức câu không đi đủ sâu vào chi tiết để giải quyết vấn đề trên.

Phân loại ý kiến ở mức khía cạnh

- Aspect-based opinion mining, tìm mọi $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ trong tài liệu cho trước d.
  - Đòi hỏi khả năng xử lý ngôn ngữ tự nhiên chuyên sâu hơn, nhưng cung cấp kết quả phong phú và ý nghĩa hơn.

- Phần này tập trung chủ yếu vào hai tác vụ trọng yếu:
  - Rút trích khía cạnh (aspect extraction)
  - Phân loại ý kiến theo khía cạnh (aspect sentiment classification)

Rút trích khía cạnh

- Rút trích các khía cạnh cần đánh giá.
- Một số ví dụ minh họa
  - “The picture quality of this camera is amazing.” → “picture quality” là khía cạnh của thực thể được biểu diễn bởi “this camera”.
  - “I do not love this camera. ” → khía cạnh GENERAL nói về chính thực thể được biểu diễn bởi “this camera”.
- Khi đề cập đến khía cạnh thì phải luôn xác định thực thể mà nó thuộc về.

Phân loại ý kiến ở mức khía cạnh

- Xác định khuynh hướng ý kiến được phát biểu trên từng khía cạnh trong câu.
  - Trong các ví dụ trước, ý kiến theo khía cạnh “picture quality” là tích cực, và ý kiến trên khía cạnh GENERAL là negative.
- Có thể áp dụng lại các phương pháp phân loại ý kiến ở mức câu và mức mệnh đề.
  - Khía cạnh nhận khuynh hướng ý kiến của câu hoặc mệnh đề.
  - Gặp khó khăn khi xử lý câu có nhiều ý kiến hỗn hợp hoặc khi cần phân tích ở mức mệnh đề/cụm từ (khó thực hiện trong môi trường văn bản không chuẩn như blog hay diễn đàn).
- **Hướng tiếp cận dựa trên lexicon** là giải pháp khả thi hơn.

Hướng tiếp cận dựa trên lexicon

- Sử dụng tập thuật ngữ ý kiến để xác định khuynh hướng ý kiến trong câu
- Giải pháp bao gồm các bước chính
  - Đánh dấu từ và cụm từ ý kiến
  - Xử lý từ chuyển đổi ý kiến
  - Xử lý mệnh đề “but”
  - Tích hợp ý kiến

Đánh dấu từ và cụm từ ý kiến

- Cho trước một câu chứa một hay nhiều khía cạnh.
- Đánh dấu mọi từ và cụm từ ý kiến trong câu.
- Gán điểm ý kiến cho (cụm) từ tích cực (+1) và tiêu cực (-1).

- Ví dụ, “The ***picture quality*** of this camera is not **great** [+1], but the ***battery life*** is long.”
  - +1 do “great” là từ ý kiến tích cực (có gì đó sai?).
  - Giả sử chưa biết từ loại ý kiến của “long” vì nó phụ thuộc nhiều vào ngữ cảnh.

Xử lý từ chuyển đổi ý kiến

- Opinion shifters hoặc shifters.
- Các từ và cụm từ có thể thay đổi khuynh hướng ý kiến.
- Từ phủ định thông dụng: not, never, none, nobody, nowhere, neither và cannot.

- *Ví dụ, "The* **picture quality** of this camera is not **great** [-1], but the ***battery life*** is long." → chuyển thành -1 do từ "not".

Xử lý từ chuyển đổi ý kiến

- Modal auxiliary verb: would, should, could, might, must ought.
  - Ví dụ, “The brake coud be improved.”
- Barely, hardly, fail, omit, neglect
  - “It works.” so với “It harly works.”, “This camera fails to impress me.”
- Lời châm biếm mỉa mai
  - Ví dụ, “What a great car, it failed to start the first day.”

- Tuy nhiên, không phải cứ xuất hiện từ chuyển đổi ý kiến là khuynh hướng ý kiến bị thay đổi.
  - Ví dụ, not only…but also.

Xử lý mệnh đề “but”

- Khuynh hướng ý kiến trước “but” và sau “but” đối lập nhau.
- Ví dụ, “The ***picture quality*** of this camera is not **great** [-1], but the battery life is long [+1].”

- Một số từ khác: “with the exception of”, “except that” và “except for”.
- Từ tương phản không phải lúc nào cũng tạo ra sự chuyển dịch ý kiến.
  - Ví dụ, “Car-x is great, but Car-y is better.”, “not only…but also”.

Tích hợp ý kiến

- Cho câu s chứa một tập khía cạnh $\{a_1, \ldots, a_m\}$ và một tập từ/cụm từ ý kiến $\{ow_1, \ldots, ow_n\}$ với điểm ý kiến đã được xác định ở những bước trước.
- Khuynh hướng ý kiến cho khía cạnh a_i trong s được xác định bởi hàm tích hợp ý kiến

$$score(a_i, s) = \sum_{ow_j \in s} \frac{ow_j.oo}{dist(ow_j, a_i)}$$

  - trong đó ow_j là từ ý kiến trong s, $dist(ow_j, a_i)$ là khoảng cách giữa khía cạnh a_i và ow_j trong s, và $ow_j.oo$ là điểm ý kiến của ow_j.

Các luật ý kiến cơ bản

- Luật ý kiến biểu diễn một khái niệm bao hàm ý kiến tích cực hoặc tiêu cực.

1. POSITIVE ::= P
2. | PO
3. | orientation shifter N
4. | orientation shifter NE
5. NEGATIVE ::= N
6. | NE
7. | orientation shifter P
8. | orientation shifter PO

  - P và PO **biểu diễn ý kiến tích cực** (positive opinion expressions).
  - N và NE **biểu diễn ý kiến tiêu cực** (negative opinion expressions).

Các luật ý kiến cơ bản

- **Từ và cụm từ ý kiến:** Bản thân (cụm) từ ý kiến có thể bao hàm ý kiến tích cực hoặc tiêu cực về khía cạnh.
  - Ví dụ, “great” trong “The picture quality is great.”

9.	P	::=	a positive opinion word or phrase.
10.	N	::=	a negative opinion word or phrase.

- **Sự kiện (không) mong muốn:** Phát biểu không chứa từ ý kiến nhưng bao hàm ý kiến tích cực/tiêu cực trong ngữ cảnh của thực thể.
  - Ví dụ, “After my wife and I slept on it for two weeks, I noticed a mountain in the middle of the mattress” $\rightarrow$ ý kiến tiêu cực

11.	P	::=	desired fact.
12.	N	::=	undesired fact.

Các luật ý kiến cơ bản

- **Định lượng cao, thấp, tăng và giảm của potential item:** Một số khía cạnh được gọi là **PPI** (positive potential item), trong khi một số khía cạnh khác được gọi là **NPI** (negative potential item).
  - Với PPI, một lượng nhỏ là tiêu cực và ngược lại là tích cực (battery life). Với NPI, một lượng nhỏ là tích cực và ngược lại là tiêu cực (hay price).
  - Ví dụ, “The battery life is short.” và “The battery life is long.” “This phone costs a lot” và “Sony reduced the price of the camera.”

```
13.    PO ::= no, low, less or decreased quantity of NPI
14.         | large, larger, or increased quantity of PPI
15.    NE ::= no, low, less, or decreased quantity of PPI
16.         | large, larger, or increased quantity of NPI
17.    NPI    ::= a negative potential item
18.    PPI    ::= a positive potential item
```

Các luật ý kiến cơ bản

- **Định lượng tăng và giảm của một hạng mục giữ ý kiến:** Việc tăng hay giảm định lượng của một hạng mục giữ ý kiến có thể làm thay đổi khuynh hướng ý kiến.
  - Hạng mục giữ ý kiến (N và P) thường là danh từ hoặc cụm danh từ.
  - Ví dụ, “This drug reduced my pain significantly.” → “pain” là từ giữ ý kiến tiêu cực, việc giảm “pain’ là hiệu ứng mong muốn, do đó đây là ý kiến tích cực về “drug”.

```
19.    PO ::= less or decreased N
20.         | more or increased P
21.    NE ::= less or decreased P
22.         | more or increased N
```

Các luật ý kiến cơ bản

- **Độ lệch khỏi tiêu chuẩn hoặc miền giá trị mong muốn:** Khía cạnh có giá trị lệch khỏi chuẩn sẽ mang ý tiêu cực.
  - Một số lĩnh vực qui định giá trị chuẩn hoặc miền giá trị mong muốn.
  - Từ ý kiến không xuất hiện trong câu.
  - Ví dụ, “This drug causes low (or high) blood pressure” và “This drug causes my blood pressure to reach 200.”

23. PO ::= within the desire value range
24. NE ::= above or below the desired value range

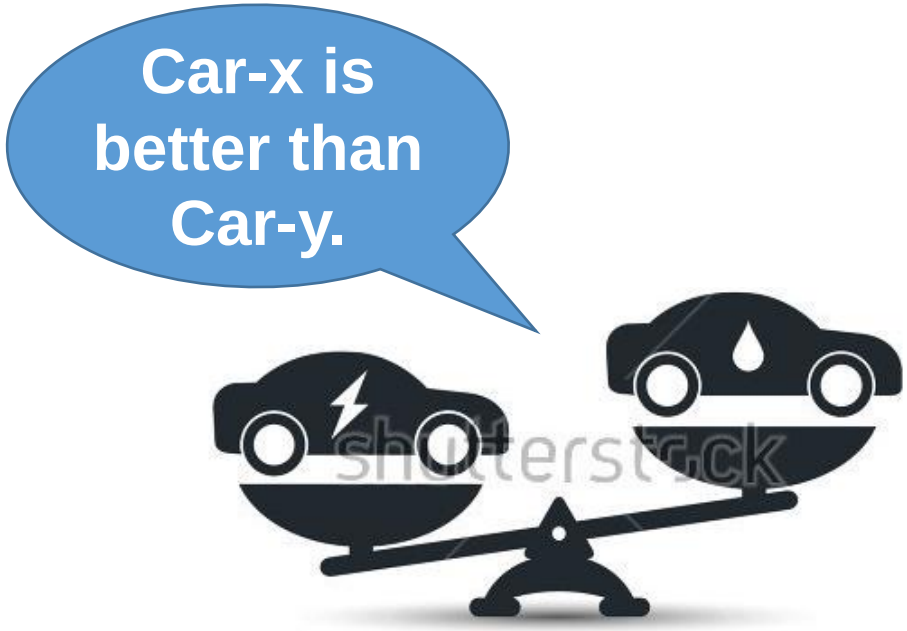
Các luật ý kiến cơ bản

- **Sản xuất và tiêu thụ tài nguyên (hoặc chất thải):** Nếu thực thể sản xuất nhiều tài nguyên thì là ý kiến tích cực. Ngược lại, nếu thực thể tiêu thụ nhiều tài nguyên, đây là ý kiến tiêu cực.
  - Ví dụ, “This washer uses a lot of water.” → water là tài nguyên → ý kiến tiêu cực về washer.
  - Tương tự, nếu thực thể sản xuất nhiều chất thải thì đây là ý kiến tiêu cực, và ngược lại. Ví dụ, “This vehicle emits too much smoke.”

```
25.      PO ::= produce a large quantity of or more resource
26.                | produce no, little or less waste
27.                | consume no, little or less resource
28.                | consume a large quantity of or more waste
         NE ::= ::= produce no, little or less resource
29.                | produce some or more waste
30.                | consume a large quantity of or more resource
31.                | consume no, little or less waste
```

Nhận xét về luật ý kiến

- Các luật có thể được biểu diễn theo nhiều cách.
  - Sử dụng từ/cụm từ khác nhau trong câu thực tế.
  - Thể hiện bằng những hình thức khác nhau trong nhiều lĩnh vực.
- Tập luật không cố định, có thể thêm/hiệu chỉnh các luật
- Câu khiến cho luật thỏa không có nghĩa là nó thực sự diễn đạt ý kiến.
  - Ví dụ, “I want a reliable car.” có từ ý kiến tích cực “reliable” nhưng câu không đưa ra ý kiến gì.



Khai thác ý kiến so sánh

So sánh

- Các **so sánh** (comparision) có ngữ nghĩa và cú pháp khác biệt so với ý kiến thông thường.
  - “The sound quality of Phone-x is better than that of Phone-y.” so với “The sound quality of this phone is great.”
- Quan hệ so sánh được chia thành bốn thể loại chính:
  - So sánh không tương đương (non-equal gradable comparison)
  - So sánh tương đương (equative comparison)
  - So sánh nhất (superlative comparison)
  - So sánh không thể xếp hạng (nongradable comparison)
- Các thể loại đầu là so sánh có thể xếp hạng, trong khi loại cuối cùng không thể xếp hạng.

Các thể loại ý kiến so sánh

- So sánh không tương đương: quan hệ *greater* hoặc *less than*, thể hiện thứ tự của các thực thể theo một số khía cạnh chung.
  - Ví dụ, “The Intel chip is faster than that of AMD.”
  - Quan hệ này còn bao gồm cả sự ưu tiên của người dùng, ví dụ, “I prefer Intel to AMD.”
- So sánh tương đương: quan hệ diễn tả hai hay nhiều thực thể tương đương nhau về một số khía cạnh chung.
  - Ví dụ, “The performance of Car-x is about the same as that of Car-y.”
- So sánh nhất: quan hệ *greater* hay *less than all others*, xếp hạng một thực thể trên tất cả những thực thể khác.
  - Ví dụ, “The Intel chip is the fastest.”

So sánh không thể xếp hạng

- Quan hệ so sánh các khía cạnh của hai hay nhiều thực thể nhưng không xếp hạng chúng.
- Thực thể A giống hoặc khác thực thể B về một số khía cạnh.
  - Ví dụ, “Coke tastes differently from Pepsi.”.
- Thực thể A có khía cạnh a_1 và thực thể B có khía cạnh a_2 (a_1 và a_2 thường có thể thay thế cho nhau).
  - Ví dụ, “Desktop PCs use external speakers but laptops use internal speakers.”
- Thực thể A có khía cạnh a mà thực thể B không có.
  - Ví dụ, “Phone-x has an earphone, but Phone-y does not have.”

Khai thác ý kiến so sánh

- Cho trước tập tài liệu chứa ý kiến D, tìm trong D tất cả sextuple ý kiến so sánh có dạng $(\boldsymbol{E_1, E_2, A, PE, h, t})$
  - E_1 và E_2 là tập thực thể đang được so sánh theo một số khía cạnh chung A (các thực thể của E_1 xuất hiện trước thực thể của E_2).
  - $PE \in \{E_1, E_2\}$ là tập thực thể được người cho ý kiến h ưu tiên hơn.
  - t là thời điểm đưa ra ý kiến so sánh.

- Ví dụ, "Canon's optics is better than those of Sony and Nikon," written by John in 2010.
  - ({Canon}. {Sony, Nikon}, {optics}, {Canon}, John, 2010)

Nhận diện câu so sánh

Đặc điểm của câu so sánh

- Hầu hết câu so sánh đều chứa tính từ/trạng từ so sánh.
  - Jindal, N. and B. Liu. *Identifying comparative sentences in text documents*. In Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006), 2006.
  - Thực nghiệm sử dụng 83 từ khóa, có thể nhận diện 98% câu so sánh (recall) với độ chính xác (precision) là 32%.
- Nhiều câu có từ so sánh nhưng không phải là câu so sánh.
  - Ví dụ, “I cannot agree with you more.”
- Nhiều câu không chứa từ so sánh nhưng lại là câu so sánh.
  - Thường là quan hệ không thể xếp hạng.
  - Ví dụ, “Cellphone-x has Bluetooth, but Cellphone-y does not have.”

Phương pháp nhận diện câu so sánh

- Các (cụm) từ khóa được sử dụng để lọc bỏ những câu không phải câu so sánh
  - Tính từ/trạng từ so sánh tương đương (JJR, RBR), ví dụ *more, less, better*, và từ kết thúc bằng *–er.*
  - Tính từ/trạng từ so sánh nhất (JJS, RBS), ví dụ, most, least, best, và từ kết thúc bằng *–est.*
  - Từ chỉ thị khác: same, similar, differ, as well as, favor, beat, win, exceed, outperform, prefer, ahead, than, superior, inferior, number one, up against, v.v.

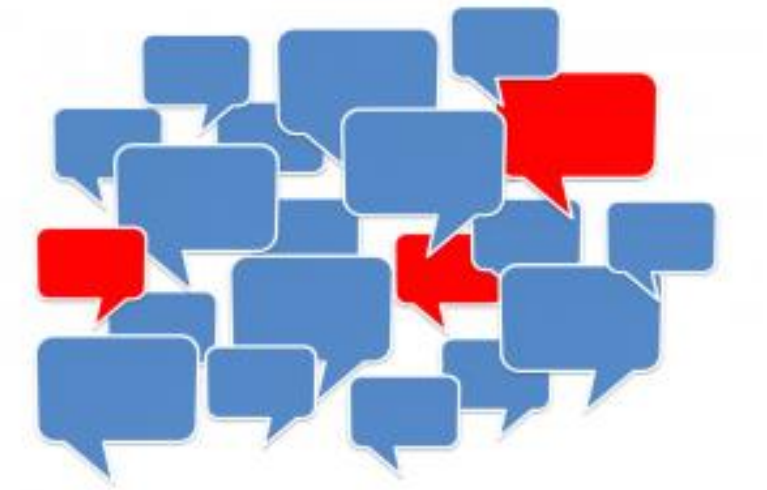
Phương pháp nhận diện câu so sánh

- Phát hiện hình mẫu từ so sánh phổ biến bằng kỹ thuật khai thác class sequential rule (CSR).
  - Mỗi luật CSR là bộ (s_i, y_i), trong đó s_i gồm các từ gần từ so sánh và $y_i \in \{comparative, noncomparative\}$ là nhãn lớp.
- Xây dựng mô hình phân lớp Bayesian với đặc trưng là vế trái của các CSR có xác suất điều kiện cao.

- Tiếp tục phân loại câu so sánh thành các loại {non-equal gradable, equative, superlative, non-gradable}.
  - Học bằng SVM với đặc trưng là từ/cụm từ.

Nhận diện thực thể ưu tiên

Tham khảo mục 11.6.3, trang 517 của tài liệu tham khảo.



Opinion Spam Detection

Sự cần thiết của ý kiến trực tuyến

Customer Reviews

Cake Pops: Tips, Tricks, and Recipes for More Than 40 Irresistible Mini Treats

177 Reviews

5 star: (142)
4 star: (21)
3 star: (9)
2 star: (3)
1 star: (2)

Average Customer Review
(177 customer reviews)

Share your thoughts with other customers
Create your own review

Search Customer Reviews
GO
Only search this product's reviews

The most helpful favorable review	The most helpful critical review
85 of 89 people found the following review helpful:	16 of 19 people found the following review helpful:
I have two words for this book, LOVE IT	**Cake Pops - Bakerella**
This has got to be the best little treat book out ever!! These cute little pops would be perfect to make for a child as well as adults. The recipe is so easy. If you don't own this book you are missing out. I plan on making these with my granddaughter. A batch of these would be perfect to give as a gift. I cannot say enough about this book. Thanks so much for writing...	I follow the Bakerella blog sit. Love the Blog... its creative and innovative. I was very excited to hear of their new book and pre-ordered it to ensure I received it right away. After receiving and reviewing the book, I have to admit I was a bit disappointed. I was expecting to see new ideas and projects however a good portion of them were duplicates from the...
Read the full review ›	Read the full review ›
Published 10 months ago by Dawna L.	Published 9 months ago by Heather R. Dugan
› See more 5 star, 4 star reviews	› See more 3 star, 2 star, 1 star reviews

Vs.

‹ Previous | 1 2 ... 18 | Next ›
Most Helpful First | **Newest First**

- Người dùng thường tìm đọc những bình luận sản phẩm trên Web với nhiều mục đích khác nhau.
  - Ví dụ, tham khảo ý kiến của người đã sử dụng sản phẩm tại các trang bán hàng (ví dụ, amazon.com) trước khi mua sản phẩm

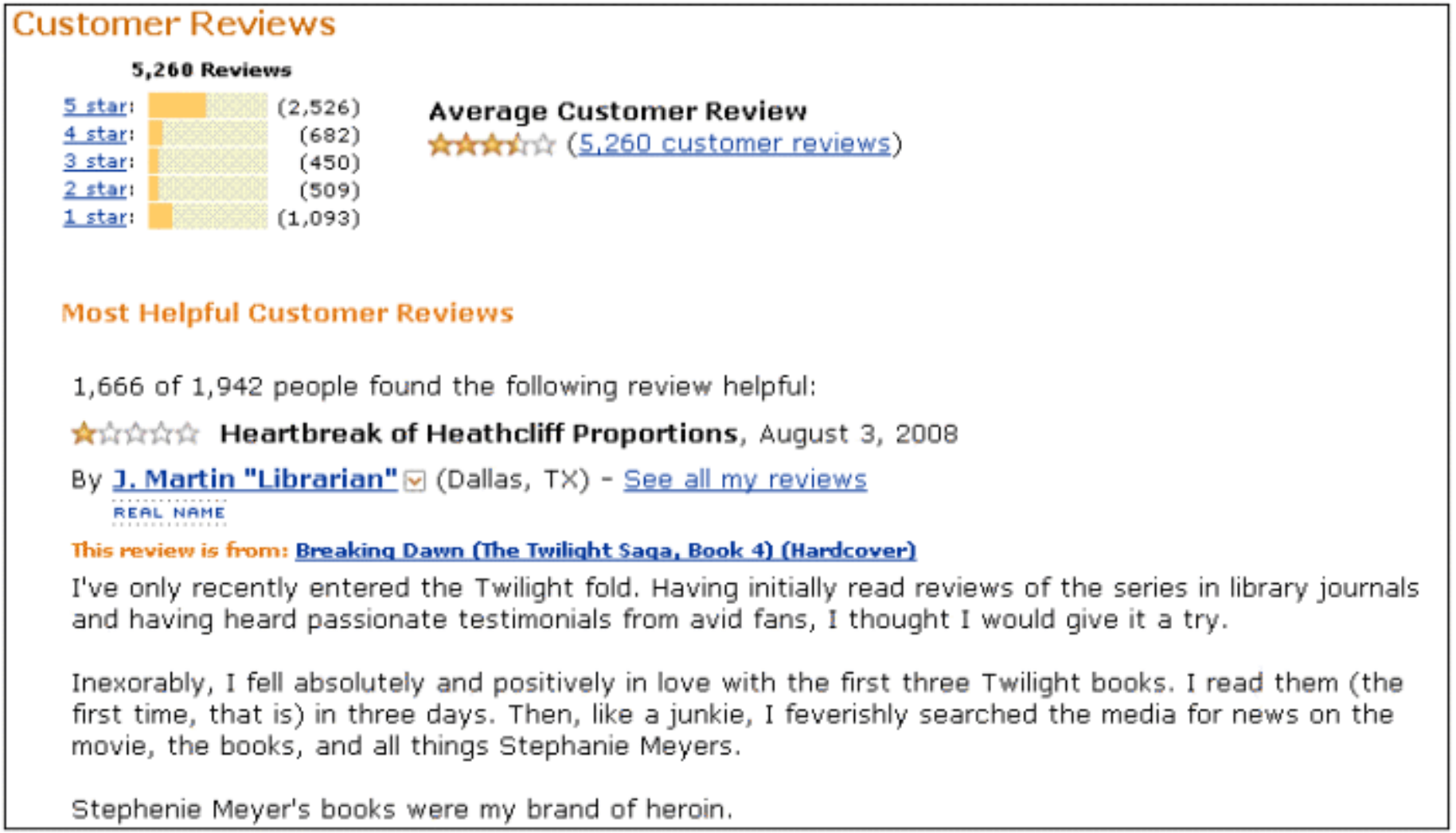
Sự cần thiết của ý kiến trực tuyến

- Các ý kiến tích cực có thể đem lại lợi ích về danh tiếng và tài chính cho các tổ chức và cá nhân.
- Ý kiến trực tuyến ngày càng được sử dụng rộng rãi trong thực tế.

Opinon spamming

- Hành động có chủ tâm (viết bình luận spam) nhằm mục đích lừa dối người đọc hoặc hệ thống khai thác ý kiến tự động.
- Các ý kiến tích cực (nhưng không xứng đáng) về một số thực thể mục tiêu để quảng cáo thực thể.
- Ý kiến tiêu cực giả hoặc có tính bất công về một số thực thể để làm tổn hại danh tiếng của thực thể.
- Từ khóa: fake opinion, bogus opinion, hoặc fake review.

Ví dụ về Opinion spam



Customer Reviews

5,260 Reviews

5 star: (2,526)
4 star: (682)
3 star: (450)
2 star: (509)
1 star: (1,093)

Average Customer Review
(5,260 customer reviews)

Most Helpful Customer Reviews

1,666 of 1,942 people found the following review helpful:

Heartbreak of Heathcliff Proportions, August 3, 2008

By J. Martin "Librarian" (Dallas, TX) - See all my reviews
REAL NAME

This review is from: Breaking Dawn (The Twilight Saga, Book 4) (Hardcover)

I've only recently entered the Twilight fold. Having initially read reviews of the series in library journals and having heard passionate testimonials from avid fans, I thought I would give it a try.

Inexorably, I fell absolutely and positively in love with the first three Twilight books. I read them (the first time, that is) in three days. Then, like a junkie, I feverishly searched the media for news on the movie, the books, and all things Stephanie Meyers.

Stephenie Meyer's books were my brand of heroin.

Opinon spam và Web spam

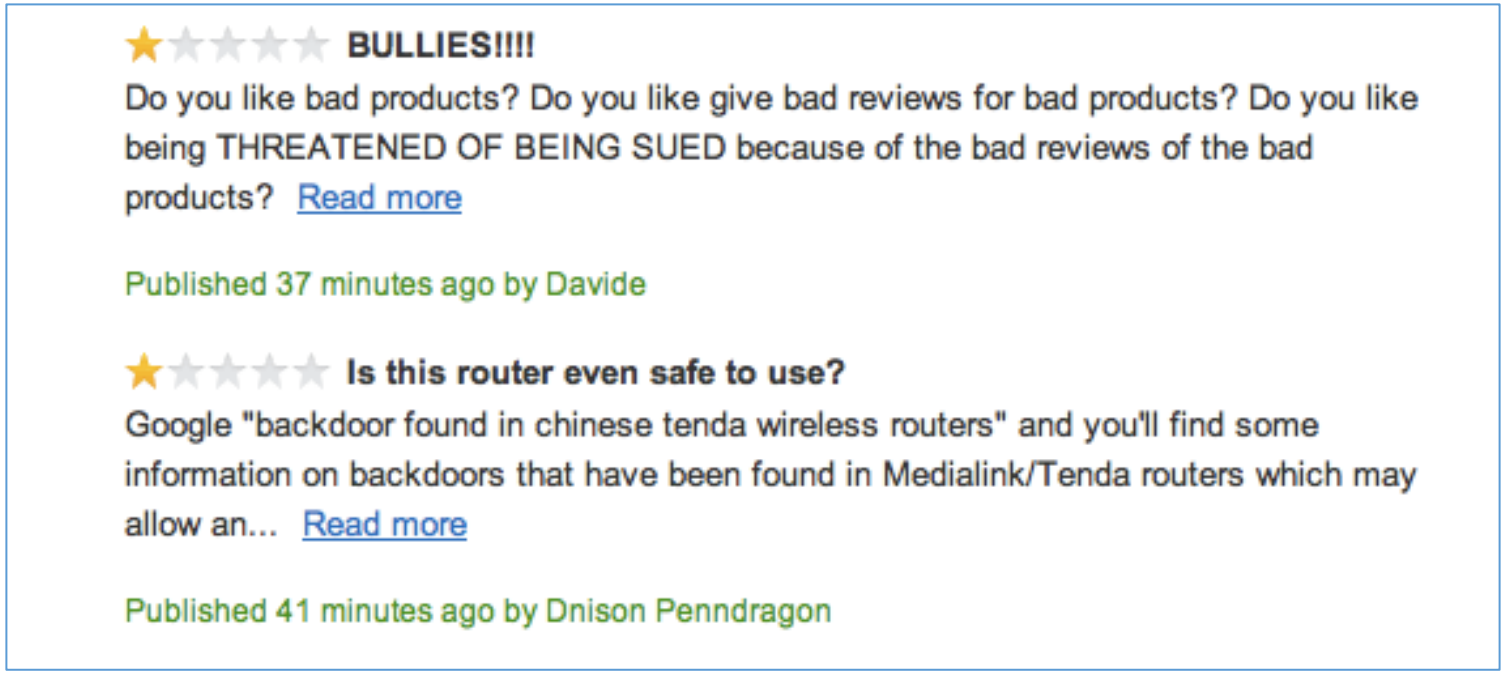
- Opinon spam rất khác so với Web spam về mặt bản chất.
- Link spam hầu như không xuất hiện trong bình luận vì thường không tồn tại liên kết giữa các bình luận.
- Content spam cũng khó xuất hiện trong bình luận theo kiểu thêm nhiều từ liên quan như trong ngữ cảnh Web.

Thể loại bình luận spam

- Loại 1 (bình luận giả): cho ý kiến tích cực không xứng đáng để quảng cáo sản phẩm hoặc cho ý kiến tiêu cực để hạ bệ danh tiếng của sản phẩm.
- Loại 2 (chỉ bình luận về nhãn hàng): không nhận xét về sản phẩm mà chỉ nhận xét về nhãn hàng, nhà sản xuất hoặc nhà cung cấp sản phẩm.
  - Mặc dù nội dung có thể hữu dụng nhưng vẫn bị xem là spam vì bình luận không hướng tới sản phẩm và thường có định kiến cao.
  - Ví dụ, trong mục bình luận về máy in HP, "I hate HP. I never buy any of their products."

Thể loại bình luận spam

- Loại 3 (không phải bình luận): Văn bản xuất hiện như là bình luận nhưng nội dung không chứa bình luận hoặc ý kiến.
  - Quảng cáo hoặc các văn bản không chứa ý kiến (ví dụ, câu hỏi đáp, nội dung ngẫu nhiên).



Nhận diện bình luận spam nguy hiểm

- Bình luận spam loại 2 và 3 thường hiếm gặp và dễ dàng bị phát hiện.
- Bình luận spam loại 1 được phân tích như sau

	Positive spam review	Negative spam review
Good quality product	1	***2***
Bad quality product	***3***	4
Average quality product	***5***	***6***

  - Sản phẩm được cho là tốt, xấu, hay trung bình tùy thuộc vào điểm số đánh giá trung bình của sản phẩm.
  - Bình luận spam trong vùng 1 và 4 không thật sự tổn hại sản phẩm,
  - Bình luận trong vùng 2, 3, 5 và 6 **cực kỳ nguy hiểm**.

Thể loại người bình luận spam

- Người spam có thể hành động theo cá nhân (ví dụ, tác giả của một cuốn sách) hoặc là thành viên của một nhóm (ví dụ, nhóm nhân viên của công ty).
- Người bình luận spam cá nhân: không cộng tác với ai khác.
  - Đăng ký như là một người dùng đơn lẻ tại (các) trang bình luận, hoặc tạo nhiều tài khoản giả với những user-id khác nhau.
- Nhóm người bình luận spam: nhiều người cùng làm việc
  - Quảng cáo sản phẩm hay làm tổn hại sản phẩm khác, đăng ký tại nhiều trang bình luận.
  - Cực kỳ nguy hiểm, có thể khống chế khuynh hướng ý kiến về sản phẩm và đánh lừa hoàn toàn người dùng tiềm năng.

Kỹ thuật che giấu cho cá nhân

- Một số trang xếp hạng người bình luận dựa vào số lượt đánh giá hữu ích của người đọc (ví dụ, amazon.com)
- Một số hệ thống cho phép người dùng gán điểm tin cậy cho người bình luận.

- Đầu tiên, spammer xây dựng danh tiếng để trở thành người bình luận đáng tin cậy bằng cách bình luận các sản phẩm không quan tâm khác một cách hợp lý.
- Sau đó, viết bình luận spam về sản phẩm mục tiêu.

Kỹ thuật che giấu cho cá nhân

- Spammer đăng ký nhiều tài khoản với user-id khác nhau tại một trang để viết bình luận spam để các bình luận không bị xem là điểm bất thường.
- Có thể sử dụng nhiều máy khác nhau để tránh bị phát hiện khi so sánh địa chỉ IP của người bình luận trong server logs.

- Chỉ viết bình luận tích cực về sản phẩm của họ **hoặc** chỉ bình luận tiêu cực về sản phẩm đối thủ.
  - Nhằm tránh kỹ thuật phát hiện spam bằng việc so sánh các bình luận của một người về sản phẩm cạnh tranh của nhãn hàng khác.

Kỹ thuật che giấu cho nhóm

- Mọi thành viên bình luận về cùng một sản phẩm để giảm độ lệch của điểm đánh giá.
- Mọi thành viên viết bình luận ngay tại thời điểm sản phẩm được giới thiệu để điều khiển khuynh hướng ý kiến.
  - Không viết nhiều bình luận spam cùng một lúc sau khi đã có nhiều bình luận khác, vì điều này tạo ra đỉnh giá trị dễ nhận thấy.
- Bình luận tại các thời điểm khác nhau để giấu đỉnh giá trị
- Chia thành các nhóm con nếu số thành viên đủ nhiều, mỗi nhóm spam ở các trang khác nhau thay vì cùng một trang.
  - Tránh bị phát hiện bởi phương pháp so sánh sự tương tự về điểm đánh giá và nội dung bình luận từ các trang khác.

Bài tập 1: Khuynh hướng ý kiến

- Một khách hàng nhận xét về sản phẩm Nokia 6610 như bên dưới

"This phone is good with a huge array of features built into it. I purchased the phone last week and have been using till then. I haven't had any problem till now. The design is sleek and the color screen has good resolution. It is very light weight and has a good signal strength. However, the main problem that I think is the with the sound quality. It is not as good as the Samsung phones that I have used earlier. When talking the voice is not very clear. But, I would definitely recommend this phone. Go for it ..."

- Xác định NĂM khía cạnh sản phẩm được đề cập trong nhận xét trên. Mỗi khía cạnh gồm tên khía cạnh A và khuynh hướng OO, kèm theo từ chỉ thị Words.

A					
OO					
Words					

Bài tập 2: Ý kiến so sánh

- Nhận xét về ba trình duyệt, Firefox, Internet Explorer và Opera như sau.

"Firefox is not faster than Internet Explorer, except for scripting, but for standards support, security and features, it is a better choice. Firefox is popular for IT users, while Internet Explorer is common for regular users. However, it is still not as fast as Opera, and Opera also offers a high level of standards support, security and features. On overall, Opera seems to be the fastest browser for Windows."

- Hãy xác định BỐN quan hệ so sánh có trong nhận xét trên.
  - Quan hệ thuộc loại nào trong bốn thể loại quan hệ so sánh chính?
  - Xác định cặp đối tượng được so sánh, E1 và E2. Quy ước: E1 là thực thể xuất hiện trước trong câu, nếu là quan hệ so sánh nhất thì E1 là thực thể ưu tiên nhất và E2 là các thực thể còn lại.
  - Khía cạnh chung được so sánh A.
  - Cụm từ chỉ thị để nhận biết quan hệ so sánh.
- Lưu ý: Câu có thể chứa nhiều hơn một quan hệ.

Bài tập 3: Khuynh hướng ý kiến

- Một khách hàng nhận xét về sản phẩm Canon-G3 như bên dưới

“The camera was a marvel. The macro works great for medical photographs and the auto mode is terrific for point and shoot. Camera powers on and ready in about a second, while offering a great speed when taking multiple pictures. Almost no lag between button activation and shutter. The quality, detail and clarity is much better than the Sony in my opinion. The viewfinder is slightly blocked by the lens. It has significantly more noise at iso 100 than the nikon 4500. The lcd screen is a little too small. Not only is it inconvenient, but also the battery life span is short.”

- Xác định NĂM khía cạnh sản phẩm được đề cập trong nhận xét trên.

Bài tập 4: Ý kiến so sánh

- Nhận xét về Instagram, Facebook, Twitter và Snapchat như sau.

"Twitter has the largest penetration potentiality as it is spreading slowly and steadily. Snapchat is one of the fastest-growing social networks, with over 100 million daily active users and 400 million snaps per day. Meanwhile, Facebook has a larger opportunity to communicate with consumers in an effortless way. Facebook receives more active users than Instagram, while Instagram is a better platform to use hashtags and post picture.. Facebook uses Facebook Live streaming feature whereas Twitter acquires Periscope which allows adding a 'go live' button. The cost of ads on Instagram is little higher than the cost of Twitter and Facebook advertising. In general, all four platforms are equally important."

- Hãy xác định NĂM quan hệ so sánh có trong nhận xét trên.

Tài liệu tham khảo

- Bing Liu. 2007. *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data.* Springer Series on Data-Centric Systems and Applications. **Chapter 11**.

Bài tập 1: Khuynh hướng ý kiến

- *GENERAL – Positive* – good
- *Design – Positive* – sleek
- *Color screen resolution – Positive* – good
- *Weight – Positive* – very light
- *Signal strength – Positive* – good
- *Sound quality – Negative* – not very clear
- Lưu ý: không sử dụng câu “It is not as good as the Samsung phones that I have used earlier.“ để dẫn chứng cho sound quality vì đây là cấu trúc so sánh.

Bài tập 2: Ý kiến so sánh

- Quan hệ so sánh không tương đương
  - E1: Firefox
  - E2: Internet Explorer
  - A: Performance speed
  - Từ chỉ thị: not faster than

  - E1: Firefox
  - E2: Internet Explorer
  - A: Standards support, security and features
  - Từ chỉ thị: a better choice

  - E1: Firefox
  - E2: Internet Explorer
  - A: Scripting
  - Từ chỉ thị: not faster than, except for

  - E1: Firefox
  - E2: Opera
  - A: Performance speed
  - Từ chỉ thị: still not as fast as

Bài tập 2: Ý kiến so sánh

- Quan hệ so sánh không thể xếp hạng
  - E1: Firefox
  - E2: Internet Explorer
  - A: types of users
  - Từ chỉ thị: popular for IT users, common for regular users
- Quan hệ so sánh nhất
  - E1: Opera
  - E2: Firefox, Internet Explorer
  - A: Brower speed (hay performance speed như các quan hệ trên)
  - Từ chỉ thị: seems to be the fastest browser