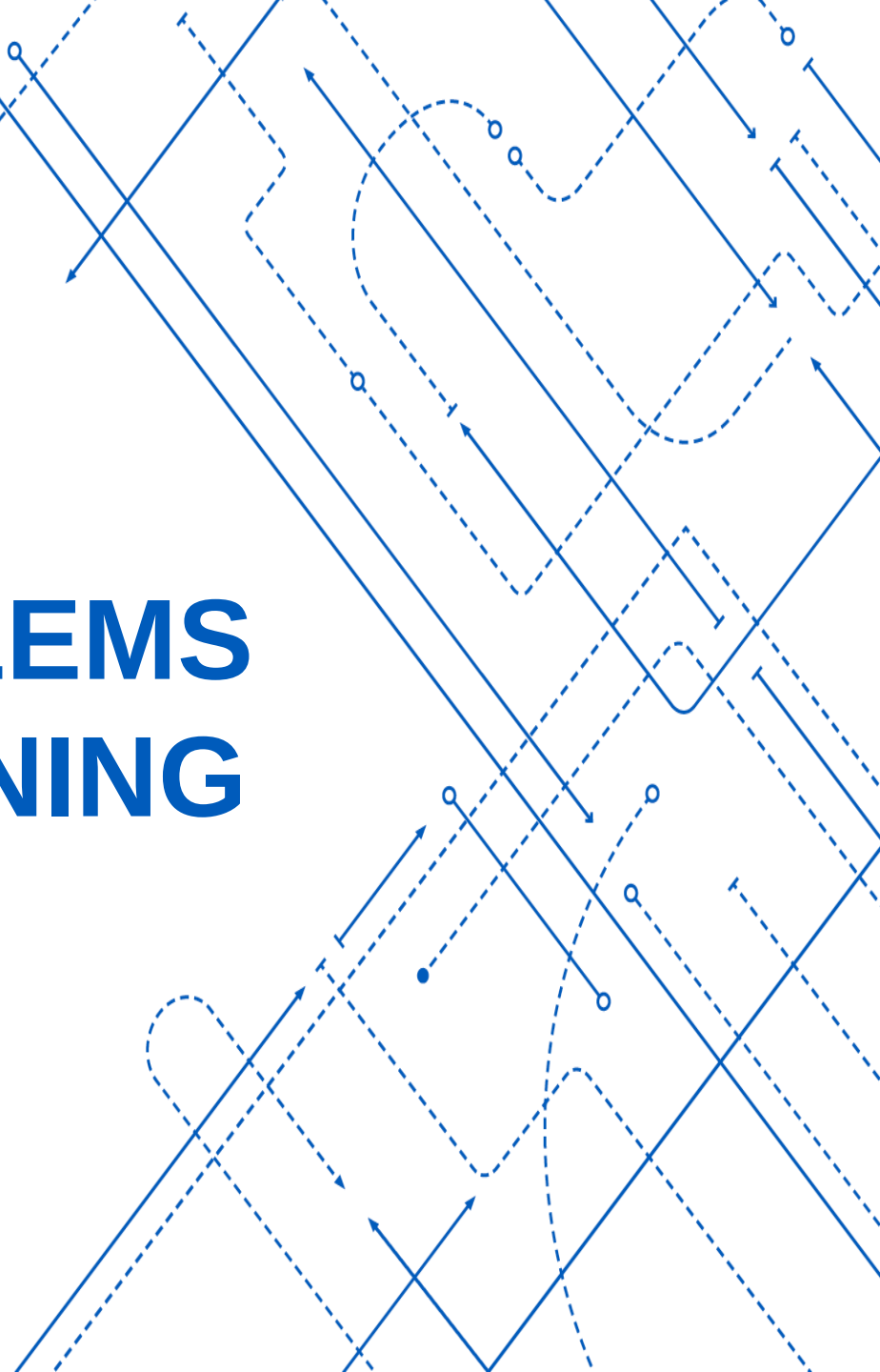
Sentiment analysis

OTHER PROBLEMS IN OPINION MINING

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn



Content outline

- Xây dựng tập thuật ngữ ý kiến
- Khai thác ý kiến so sánh
- Phát hiện ý kiến spam

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.

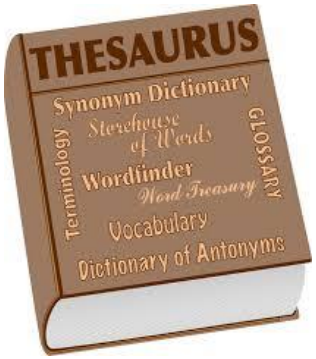
Phân loại thuật ngữ ý kiến

- **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

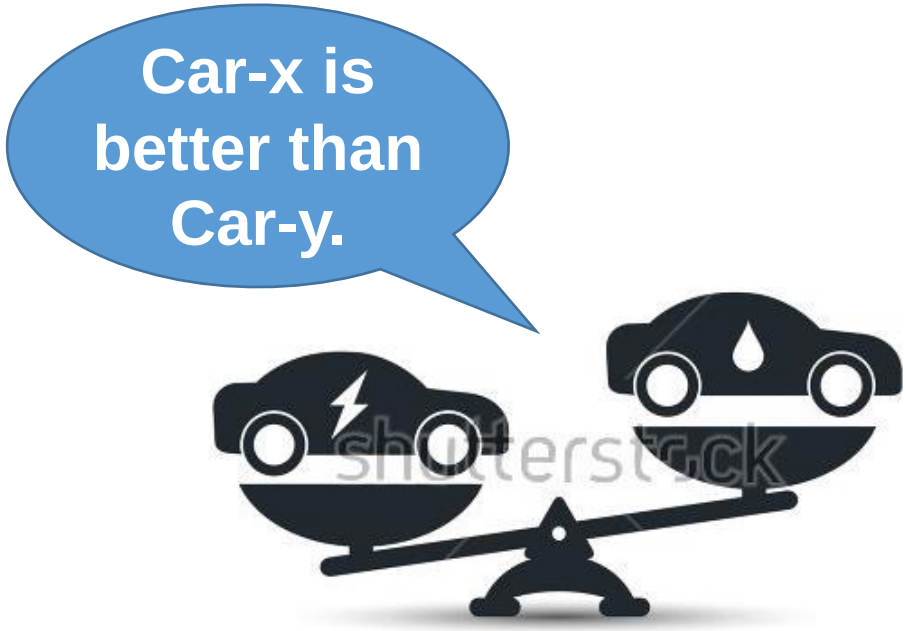
- Căn cứ vào mẫu cú pháp hoặc mẫu đồng hiện, và tập từ hạt giống ý kiến để tìm các từ khác trong kho ngữ liệu lớn.
- Sentiment consistency: nhận diện tính từ ý kiến (và khuynh hướng) khác dựa vào ràng buộc ngôn ngữ trên từ nối.
  - AND, OR, BUT, EITHER–OR, và NEITHER–NOR
  - Các tính từ liên hiệp bởi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



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.

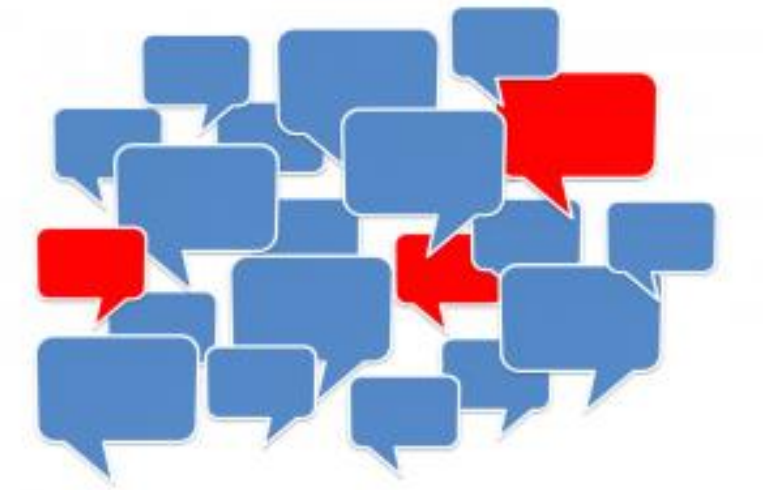
Phương pháp nhận diện câu so sánh

- 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

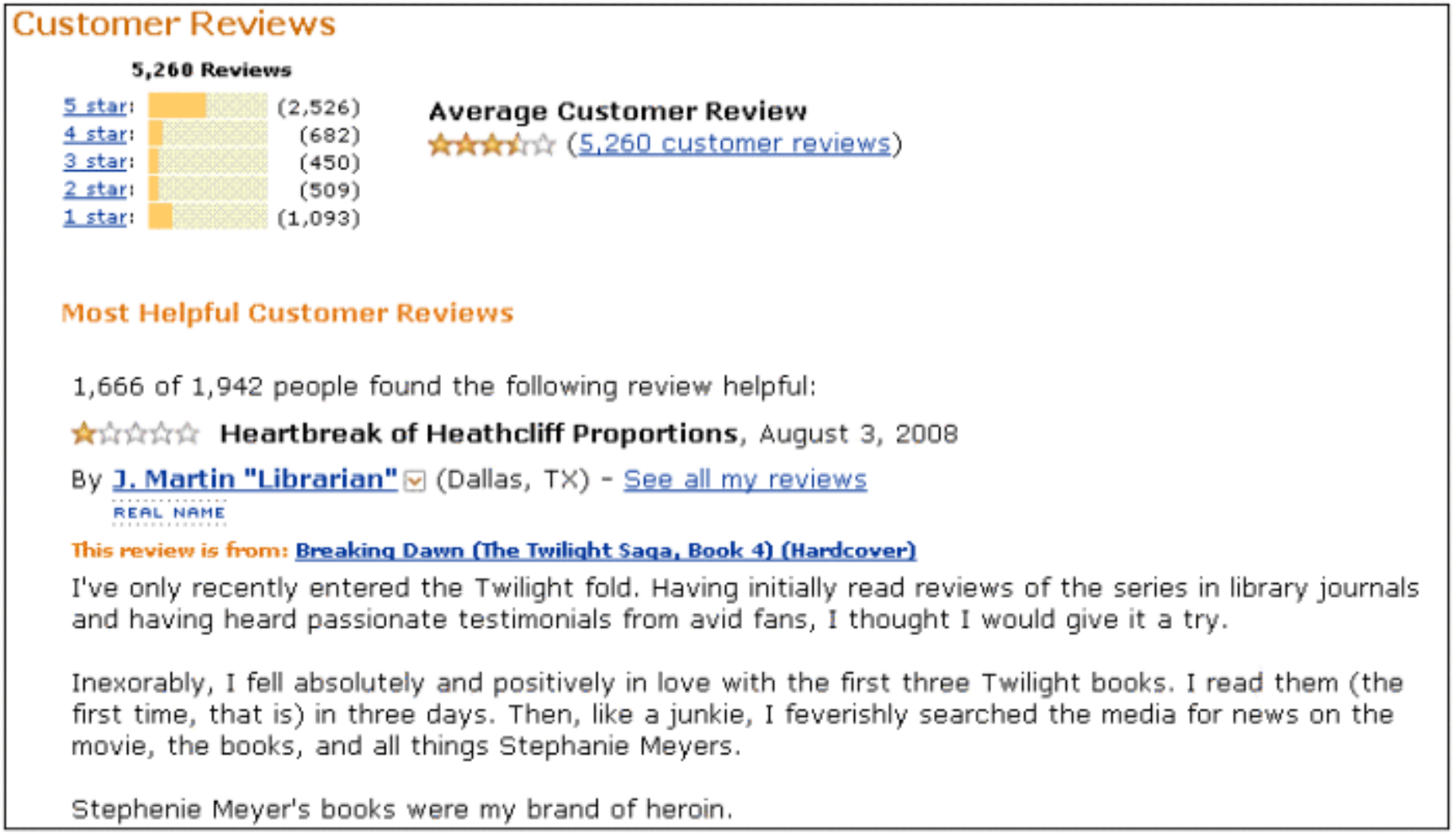
Sự cần thiết của ý kiến trực tuyến

- 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

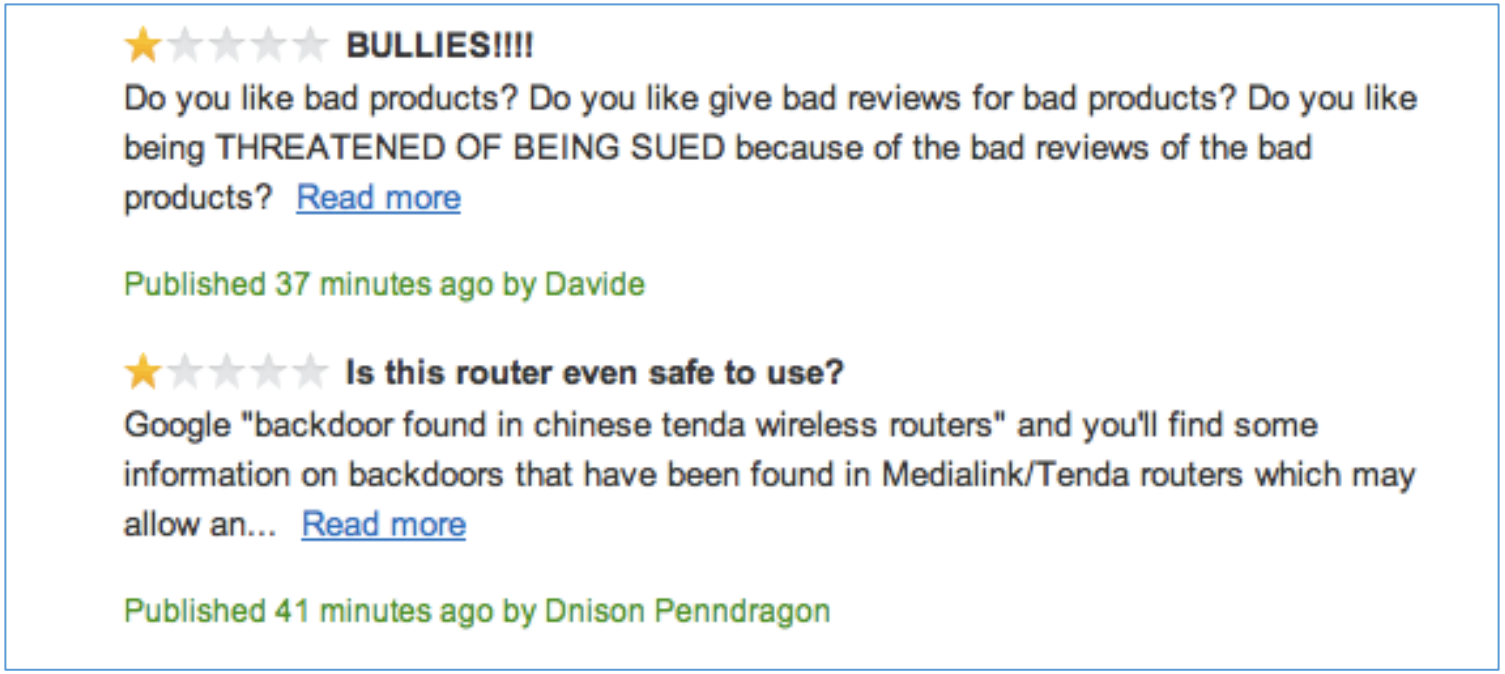
- Opinon spam rất khác so với Web spam về mặt bản chất.
- Link spam hầu như không xuất hiện trong bình luận vì thường không tồn tại liên kết giữa các bình luận.
- Content spam cũng khó xuất hiện trong bình luận theo kiểu thêm nhiều từ liên quan như trong ngữ cảnh Web.

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.

Phát hiện spam

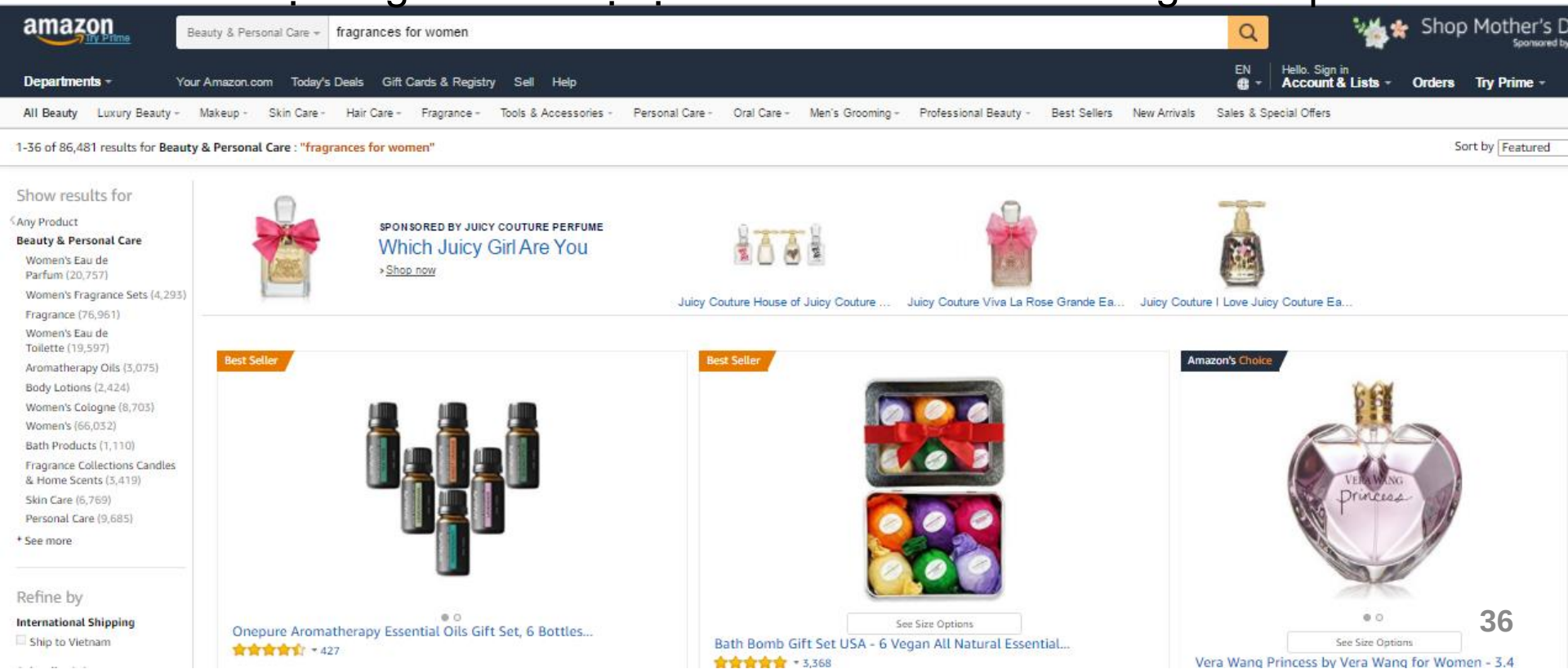
- **Phát hiện spam** (spam detection) có thể được phát biểu thành bài toán **phân lớp** với hai lớp, **spam** và **non-spam**.
- Cần có giải pháp khác nhau cho các loại spam khác nhau.
- Bình luận spam loại 2 và 3 dễ dàng bị phát hiện bằng các kỹ thuật học phân lớp truyền thống.
  - Có thể gán nhãn thủ công để xây dựng tập huấn luyện vì dễ nhận ra những bình luận spam dạng này.

Phát hiện bình luận spam loại 2 & 3

- **Đặc trưng lấy bình luận làm trung tâm.**
  - Từ rút trích trong bình luận, số lần nhãn hàng được đề cập, tỉ lệ từ ý kiến, độ dài của bình luận, số phản hồi cho rằng bình luận hữu ích.
- **Đặc trưng lấy người bình luận làm trung tâm.**
  - Giá trị trung bình và độ lệch chuẩn của điểm đánh giá được cho bởi người bình luận.
  - Có bao nhiêu bình luận là bài viết đầu tiên về sản phẩm, so với tổng số bình luận mà người này đã viết.
  - Có bao nhiêu trường hợp mà người này là người bình luận duy nhất cho sản phẩm.

Phát hiện bình luận spam loại 2 & 3

- **Đặc trưng lấy sản phẩm làm trung tâm.**
  - Giá thành sản phẩm.
  - Thứ hạng bán hàng của sản phẩm căn cứ vào số lượng đã bán.
  - Giá trị trung bình và độ lệch chuẩn của điểm đánh giá sản phẩm.



Phát hiện bình luận spam loại 1

- Rất khó để nhận diện bình luận giả bằng cách đọc nội dung bình luận vì spammer ngụy trang chúng rất tinh vi.
- Gán nhãn thủ công xây dựng tập huấn luyện không khả thi.

- Bình luận trùng từ cùng userid về cùng sản phẩm.
- Bình luận trùng từ các userid khác nhau về cùng sản phẩm.
- Bình luận trùng từ cùng userid về các sản phẩm khác nhau.
- Bình luận trùng từ các userid khác nhau về các sản phẩm khác nhau.

Phát hiện bình luận spam loại 1

- Loại 1 có thể xảy ra do bất cẩn của người bình luận (nhấn nút submit nhiều lần, có thể xác định dựa vào ngày giờ) hoặc người bình luận quay trở lại viết bình luận cập nhật sau khi đã sử dụng sản phẩm một thời gian.
- Tuy nhiên, ba loại sau hầu như chắc chắn là bình luận giả.

- Loại bỏ các bình luận cho những hình thức khác nhau của cùng một sản phẩm.
  - Ví dụ, hardcover và paperback đối với sản phẩm sách.
- Xây dựng mô hình phân lớp bằng Logistic Regressions.

Phát hiện bình luận spam loại 1

- Bình luận tiêu cực ngoại biên (điểm đánh giá lệch nhiều theo chiều hướng âm so với giá trị trung bình) là hình thức spam nghiêm trọng.
  - Trong khi đó, bình luận tích cực ngoại biên không nghiêm trọng lắm.
- Những bình luận duy nhất về sản phẩm thường là spam.
- Người bình luận xếp hạng cao cũng có thể là spammer.
  - Không giống như khách hàng thông thường, những người này viết đến hàng chục ngàn bình luận và do đó khả nghi.

Phát hiện bình luận spam loại 1

- Bình luận spam có thể nhận được phản hồi hữu dụng trong khi bình luận non-spam lại nhận phản hồi kém.
  - Người đọc dễ dàng bị các bình luận spam lừa dối. Ngoài ra, con số phản hồi hữu dụng cũng có thể bị spam.
- Sản phẩm có thứ hạng bán hàng thấp thường dễ bị spam.
  - Đây là tín hiệu tốt để nhận diện spam vì hoạt động spam thường tập trung vào sản phẩm có doanh số thấp và vì khó có thể làm tổn hại danh tiếng của sản phẩm đang được ưa chuộng chỉ bằng vài bình luận spam.

Hướng tiếp cận khác học có giám sát

- Giải quyết bài toán phát hiện spam bằng học có giám sát gặp khó khăn về việc gán nhãn thủ công dữ liệu huấn luyện.
- Hướng tiếp cận khác: nhận diện hành vi **phi điển hình** của người bình luận để phát hiện spammer.
  - Ví dụ, nếu một người chỉ viết toàn bình luận tiêu cực về một nhãn hàng trong khi những người khác đều nhìn nhận tích cực thì đương nhiên người này sẽ thuộc diện tình nghi spam.

Exercises

Bài tập 1: 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 2: 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.