

Annotation Guideline

Thành viên nhóm:

- 22022517 Bùi Tiến Sâm - Nhóm trưởng
- 22022521 Nguyễn Văn Mạnh
- 22022597 Trịnh Đắc Phú

Bài toán: Phân loại bình luận trên trang bán hàng Amazon, là khen hay chê.

Mô tả:

- **Input** : Một câu bình luận (hoặc các feature khác của nó).
- **Output**: Nhãn của câu bình luận (khen, chê, không liên quan).

1. Mục đích của quá trình gán nhãn

Là để phân loại các bình luận mua hàng thành 3 loại chính:

- Bình luận **khen** sản phẩm (**positive**)
- Bình luận **chê** sản phẩm (**negative**)
- Bình luận **không liên quan** (**spam**)

nhằm mục đích tạo ra bộ dữ liệu cho việc học có giám sát.

2. Loại nhãn và định nghĩa

- Bình luận **khen** sản phẩm (**positive**): bình luận mà người viết đạt được sự hài lòng, ấn tượng tích cực về sản phẩm, thường có những từ như: “*good, well, great, fine,...*” hoặc không có ấn tượng xấu về sản phẩm.
- Bình luận **chê** sản phẩm (**negative**): bình luận trong đó người viết đạt sự không hài lòng, thất vọng hoặc gặp vấn đề với sản phẩm, những từ thường có vd: “*bad, slow, expensive, not worth, stop working...*”
- Bình luận **không liên quan** (**spam**): trong bình luận người nói không nói gì đến sản phẩm, hoặc thể hiện cảm xúc với sản phẩm,...

3. Ví dụ minh họa

- Bình luận **khen** sản phẩm:
“*The cord is sturdy and well made. Plenty long enough. Gets the job done.*”



Robert D

★★★★★ Gets the job done

Reviewed in the United States on March 12, 2024

Size: 10ft | Color: Black | **Verified Purchase**

The cord is sturdy and well made. Plenty long enough. Gets the job done.

Helpful

Report

- Bình luận **chê** sản phẩm:

“Totally stopped working after 2 weeks. Not really worth the effort to return it.”



Sarah

☆☆☆☆☆ Lasted 2 weeks

Reviewed in the United States on March 10, 2024

Size: 10ft | Color: Black | **Verified Purchase**

Totally stopped working after 2 weeks. Not really worth the effort to return it.

Helpful

Report

- Bình luận **không** liên quan:

“My boyfriend do not want to kiss cuz my face looks like your anker. Change your porduct or custom must miss you guy”



Mrs. H

☆☆☆☆☆ Kiss or miss

Reviewed in the United States on February 16, 2024

Size: 10ft | Color: Black | **Verified Purchase**

My boyfriend do not want to kiss cuz my face looks like your anker. Change your porduct or custom must miss you guy

One person found this helpful

Helpful

Report

4. Quy trình gán nhãn

- Tập dữ liệu sẽ được chia ra làm 3 phần, sẽ có ba người gán nhãn mỗi người sẽ gán nhãn 2 phần chéo nhau (ví dụ: người thứ nhất gán phần 1 và 2 , người thứ 2 gán phần 2 và 3, người thứ 3 gán phần 1 và 3).



- Mỗi người sẽ đọc những phản hồi của người mua để hiểu ý kiến của người viết.
- Xác định nhãn của sản phẩm là một trong ba loại nêu trên.
- Sau khi gán nhãn xong, cả nhóm sẽ kiểm tra chéo nhau xem độ chính xác giữa hai người là bao nhiêu, nếu kết quả gán nhãn không giống nhau, cần họp và quyết định những nhãn đó.

5. Đánh giá quy trình gán nhãn

- Lợi ích: giảm khối lượng công việc nhưng vẫn đem lại hiệu quả: mặc dù cả 3 người cùng gán nhãn một điểm dữ liệu, khả năng xảy ra tình huống gây bối rối vẫn cao hơn, tuy nhiên, với 2 người cùng gán nhãn một điểm dữ liệu thì tình huống gây bối rối vẫn xảy ra.
- Nhược điểm: vẫn có thể bỏ sót những điểm dữ liệu gây bối rối nếu như cả 2 người có cùng quan điểm và 2 chưa phải đủ lớn để đánh giá có gây bối rối hay không.

6. Một số vấn đề

- Nếu bình luận có cả phần khen và phần chê, thì phân loại kiểu gì?
⇒ Dựa vào số lượng ý khen và ý chê của người viết, ý nào chiếm ưu thế hơn thì bình luận sẽ gán nhãn đó.

Trong trường hợp số ý là như nhau hoặc không thể phân định được, để nhãn là **Unknown**, đợi đến cuộc họp gần nhất để thảo luận và phân định.