**Annotation guideline**

**Bài toán: Phân loại chủ đề của các bài báo trên trang NPR và NBC.**

**Input: Một bài báo gồm cả tiêu đề và nội dung**

**Output: Nhãn của bài báo (Chính trị, kinh doanh, y tế, thể thao, khoa học).**

# **Mục đích của quá trình gán nhãn**

Bài toán được đề ra là bài toán phân lớp, với cách tiếp cận là học máy có giám sát, nên cần sử dụng dữ liệu có gán nhãn.

Như đã đề cập, gán nhãn là để phân loại các bài báo thành 5 loại chính:

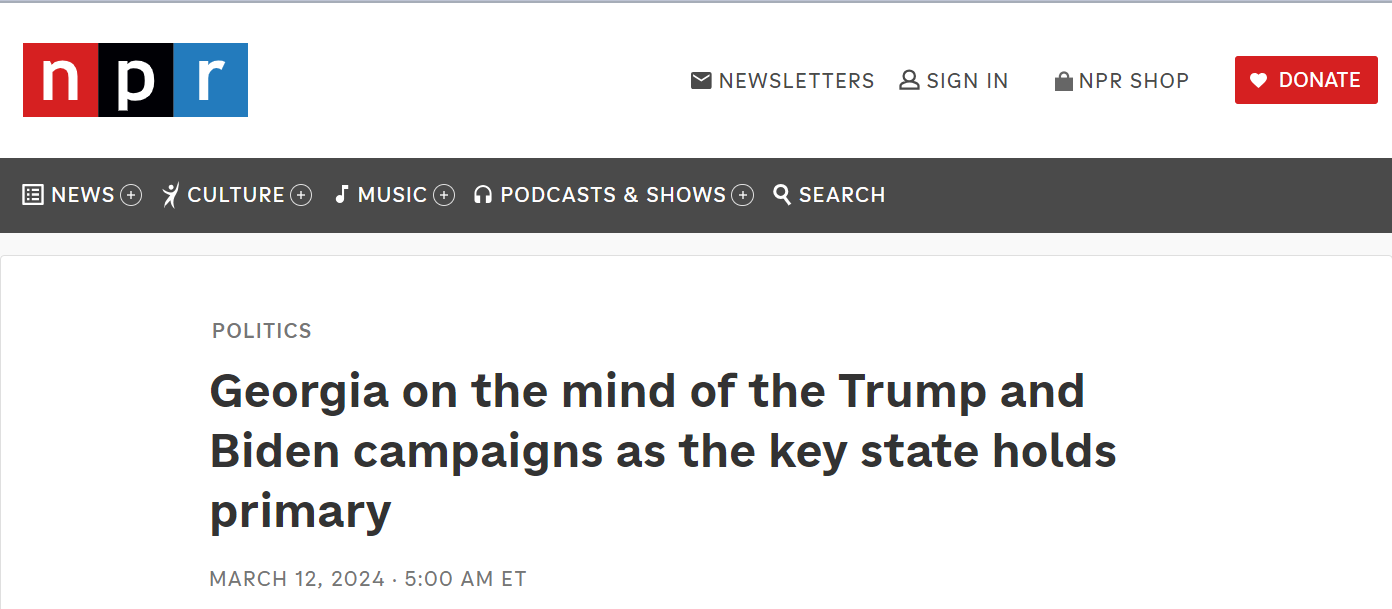
* Báo chính trị (**politics**)
* Báo kinh doanh (**business**)
* Báo y tế (**health**)
* Báo thể thao (**sports**)
* Báo khoa học (**science**)

# **Đặc trưng của các loại nhãn**

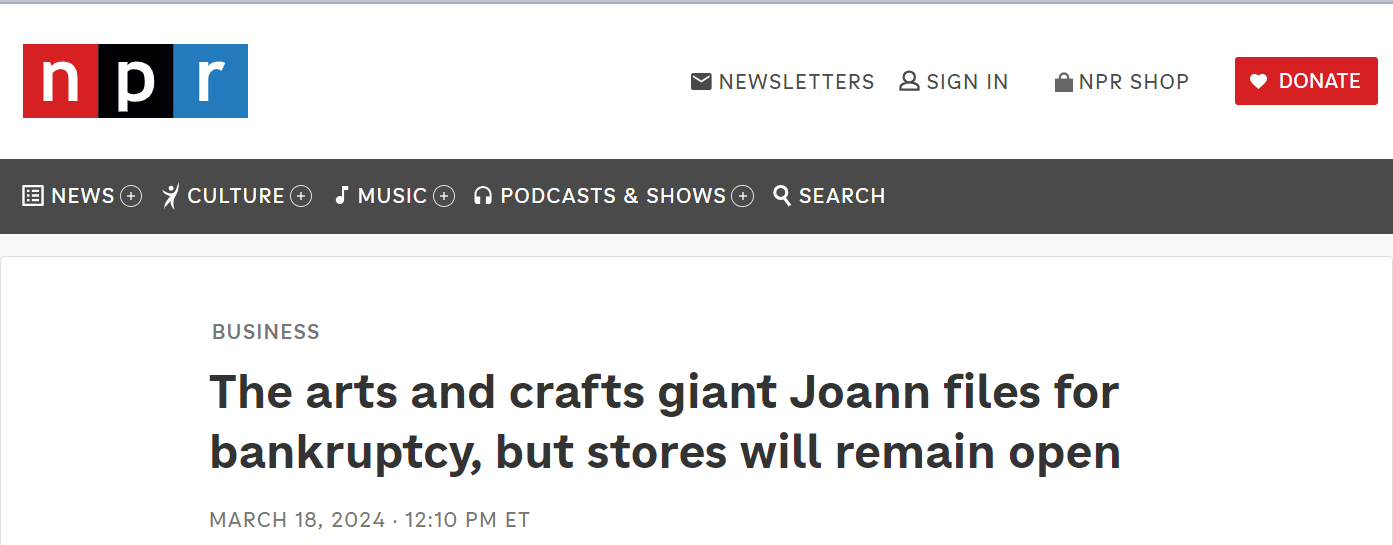
* **Báo chính trị (politics)**: Đặc trưng bởi sự xuất hiện của tên các chính trị gia và các từ ngữ đặc thù liên quan đến chính trị như “vote”, “protest”, “Trump”, “Biden”, “white house”, ...
* **Báo kinh doanh (business)**: Đặc trưng bởi sự xuất hiện của nhiều số liệu, thường liên quan đến doanh thu, lời/lỗ, cổ phần, ...
* **Báo y tế (health)**: Đặc trưng bởi các sự xuất của các từ có chủ đề y học như “hospital”, “doctor”, “nurse”, “medicine”, “treatment”, “health”, ...
* **Báo thể thao (sports)**: Đặc trưng bởi sự xuất hiện của các môn thể thao, và một số giải thể thao lớn như “Olympic”, “Super Bowl”, “All Star”, “football”, “basketball”, ...
* **Báo khoa học (science)**: Đặc trưng bởi sự xuất hiện của các từ ngữ đặc thù trong khoa học tự nhiên và xã hội như “space”, “archaeology”, “history”, ...

# **Ví dụ**

* Báo chính trị:



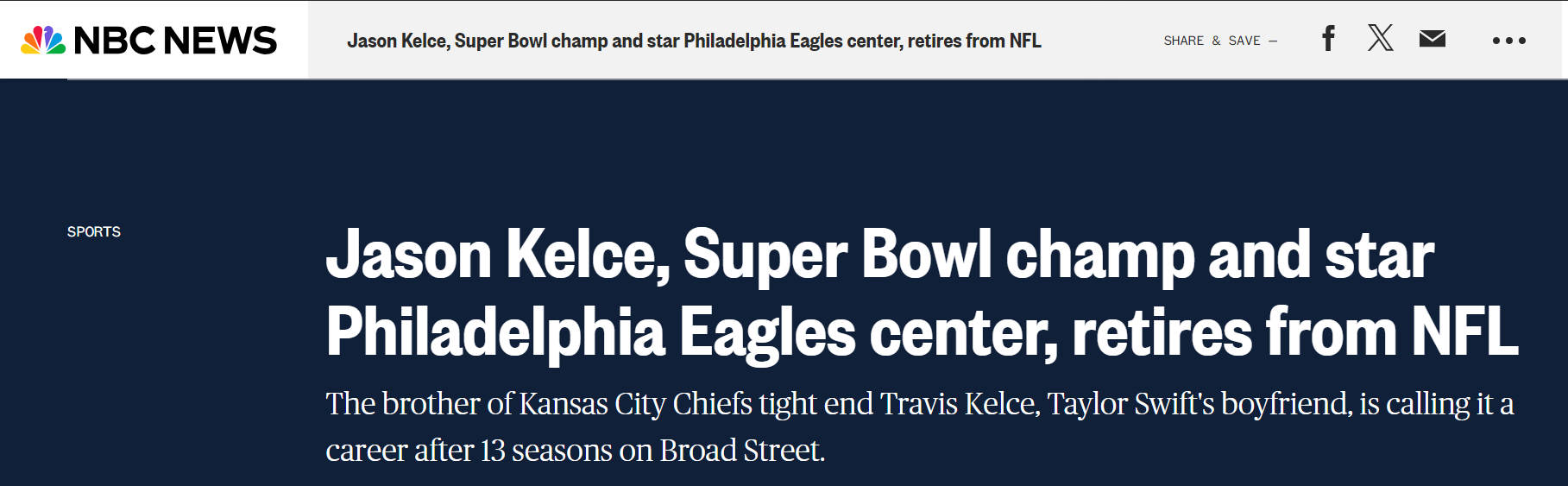
* Báo kinh doanh:



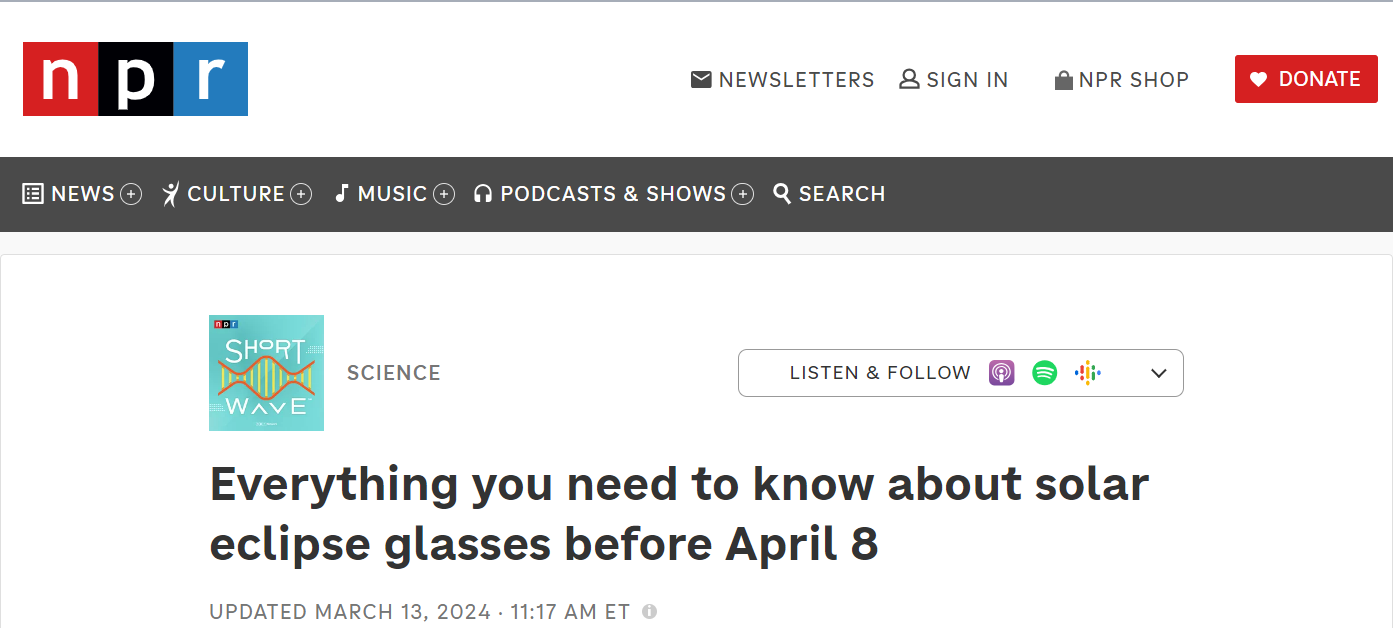
* Báo y tế:



* Báo thể thao:

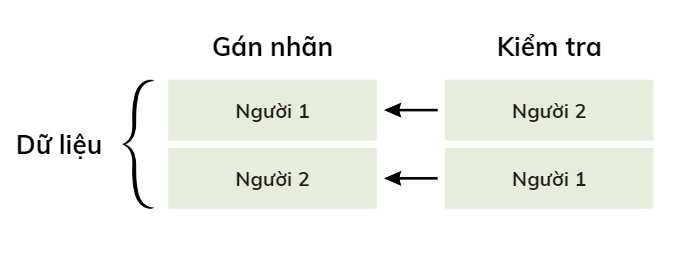


* Báo khoa học:



# **Quy trình gán nhãn**

Việc gán nhãn được thực hiện bởi 2 thành viên trong nhóm. Tập dữ liệu sẽ được xáo trộn và chia đôi cho mỗi người. Sau khi gán nhãn xong, 2 người tráo đổi tập dữ liệu và gán nhãn lần hai cho tập dữ liệu của người kia. Bất cứ bài báo nào có xung đột trong việc gán nhãn đều được lấy ra để thảo luận và thống nhất lại.



# **Đánh giá quy trình**

* Ưu điểm: Kết quả gán nhãn sau cùng sẽ trung hòa cả 2 quan điểm đã đi đến thống nhất của 2 người gán nhãn. Điều này giúp tránh khỏi cách tiếp cận 1 chiều nếu chỉ có 1 người gán nhãn.
* Nhược điểm: Quá trình phân chia công việc trong nhóm cử ra 2 trên 3 người gán nhãn, đồng nghĩa với việc quan điểm gán nhãn của người thứ 3 sẽ phải phụ thuộc vào 2 người còn lại

# **Các vấn đề gặp phải**

Các bài báo đôi khi sẽ không hoàn toàn thuộc về 1 loại chủ đề. Sự xuất hiện của những bài báo nhập nhằng gây khó dễ trong quá trình gán nhãn (các bài báo nói về kinh tế trong chính trị, chính trị trong kinh tế, y tế trong thể thao, y tế trong khoa học, khoa học trong chính trị, ...)

Ví dụ:

* Bài báo có đề cập đến “basketball” và một số từ ngữ khác liên quan đến thể thao nhưng nó lại thuộc loại báo kinh doanh



* Bài báo thuộc loại báo chính trị nhưng có tag nhỏ là liên quan đến y tế, sức khỏe

