**BÁO CÁO**

**BÀI THỰC HÀNH 2**

1. **Bài toán: Nhận diện bình luận độc hại trên mạng xã hội**
2. **Các thành viên:**

|  |  |
| --- | --- |
| MSSV | Tên |
| 21522678 | Phạm Trung Tín |
| 21522515 | Nguyễn Việt Quang |
| 21522429 | Trần Nguyễn Yến Nhi |

1. **Nội dung báo cáo:**
   1. **Các tiêu chuẩn của một benchmark**

Một benchmark trong máy học là một tập dữ liệu hoặc bài toán thử nghiệm được sử dụng để đánh giá và so sánh hiệu suất của các mô hình hoặc phương pháp khác nhau. Để một benchmark đáng tin cậy và hữu ích, nó cần phải tuân theo các tiêu chuẩn cụ thể.

Dưới đây là các tiêu chuẩn quan trọng của một benchmark trong máy học:

1. Tính đại diện (Representativeness):

Benchmark cần phản ánh một phần quan trọng của thực tế hoặc mục tiêu được nghiên cứu. Dữ liệu trong benchmark nên thể hiện một phạm vi đa dạng và phân phối tương tự với thực tế để đảm bảo tính đại diện.

Ví dụ: Một benchmark trong lĩnh vực nhận dạng hình ảnh nên chứa hình ảnh từ nhiều nguồn khác nhau, bao gồm cả hình ảnh tự nhiên và hình ảnh trong các môi trường công nghiệp. Điều này đảm bảo ràng benchmark phản ánh đa dạng của hình ảnh thực tế mà mô hình có thể gặp phải.

1. Sự đa dạng (Diversity):

Benchmark cần bao gồm sự đa dạng về trường hợp hoặc tình huống thử nghiệm. Điều này giúp đảm bảo rằng các mô hình hoặc phương pháp được thử nghiệm chứa đựng sự phức tạp cần thiết để xử lý nhiều tình huống.

Ví dụ: Một benchmark trong việc dự đoán nhà cần chứa dữ liệu về các loại nhà khác nhau, từ căn hộ chung cư đến nhà đất riêng lẻ hay biệt thự. Điều này đảm bảo rằng các mô hình được thử nghiệm có thể xử lý nhiều loại tài sản bất động sản.

1. Điểm đánh giá (Evaluation Metrics):

Benchmark cần có các tiêu chuẩn rõ ràng và thang đo để đánh giá hiệu suất. Điều này bao gồm các phân đoạn (ground truth) hoặc các tiêu chí đánh giá khác để đo lường kết quả của mô hình.

Ví dụ: Một benchmark trong phân loại hình ảnh nên định rõ các tiêu chuẩn đánh giá như tỉ lệ phân loại đúng (accuracy), độ chính xác (precision), độ phủ (recall), và F1-score để đánh giá hiệu suất mô hình.

1. Khả năng so sánh (Comparability):

Benchmark cần phù hợp để so sánh giữa các phương pháp hoặc mô hình khác nhau. Điều này có thể liên quan đến việc benchmark đã được sử dụng rộng rãi trong cộng đồng nghiên cứu hoặc có thể tạo ra khung thử nghiệm chung cho mọi người.

Ví dụ: Một benchmark trong việc dự đoán thời tiết phải chứa đủ nhiều dữ liệu từ nhiều nguồn thời tiết khác nhau để đảm bảo mô hình có độ tin cậy và đủ số lượng mẫu để dự đoán.

1. Kích thước lớn và số lượng mẫu đủ lớn (Large-Scale and Sufficient Samples):

Benchmark nên đủ lớn để cung cấp dữ liệu đáng kể để thử nghiệm hiệu suất. Số lượng mẫu cần phải đủ để đảm bảo sự độc lập và độ tin cậy của kết quả đánh giá.

Ví dụ: Một benchmark trong việc dự đoán thời tiết phải chứa đủ nhiều dữ liệu từ nhiều nguồn thời tiết khác nhau để đảm bảo mô hình có độ tin cậy và đủ số lượng mẫu để dự đoán.

1. Sự chắc chắn và độ tin cậy (Uncertainty and Reliability):

Cần kiểm tra tính chắc chắn của dữ liệu trong benchmark và đảm bảo rằng các giá trị đã được kiểm tra và làm sạch để loại bỏ lỗi.

Ví dụ: Benchmark trong việc phân loại văn bản phải kiểm tra sự đáng tin cậy của nhãn bằng cách sử dụng nhiều người đánh giá độc lập hoặc sử dụng các tiêu chuẩn quy định.

1. Khả năng tái sử dụng (Reusability):

Benchmark nên được thiết kế để có thể tái sử dụng và chia sẻ với cộng đồng nghiên cứu. Điều này giúp tạo nên một khung thử nghiệm chuẩn mực cho tất cả mọi người.

Ví dụ: Một benchmark trong việc dự đoán giá cổ phiếu cần được chia sẻ công khai với cộng đồng nghiên cứu để có thể tái sử dụng và kiểm tra bởi các nhà nghiên cứu khác.

1. Bảo mật và quyền riêng tư (Security and Privacy):

Đảm bảo rằng benchmark tuân theo các quy tắc bảo mật và quyền riêng tư, đặc biệt khi nó chứa thông tin cá nhân hoặc nhạy cảm.

Ví dụ: Benchmark chứa dữ liệu y tế như hồ sơ bệnh nhân cần tuân thủ nghiêm ngặt các quy tắc về quyền riêng tư và bảo mật để đảm bảo thông tin cá nhân không bị tiết lộ.

1. Sự độc lập và khách quan (Independence and Objectivity):

Benchmark cần phải được xây dựng và thực hiện một cách độc lập và không thiên lệch. Khách quan trong việc đánh giá và so sánh hiệu suất.

Ví dụ: Để đảm bảo sự độc lập, một benchmark trong việc dự đoán thị trường chứng khoán không nên được tài trợ bởi các công ty chứng khoán hoặc tổ chức tài chính khác để tránh thiên lệch.

1. Tài liệu hóa và hướng dẫn (Documentation and Guidelines):

Benchmark nên đi kèm với tài liệu hướng dẫn và mô tả cụ thể về cách sử dụng nó, cách đánh giá và các quy tắc về độ đồng thuận.

Ví dụ: Benchmark đi kèm với tài liệu hướng dẫn chi tiết về cách sử dụng nó, đánh giá mô hình, và cách tiếp cận các dữ liệu.

* 1. **Mô tả bài toán**

Nhận diện bình luận độc hại trên mạng xã hội, cụ thể là các bình luận của các video trên youtube(input là là các bình luận, output là 1 trong 3 nhãn HATE, OFFENSIVE và CLEAN)

* 1. **Guideline xây dựng dữ liệu**

\* Định nghĩa:

1. Bình luận sạch (CLEAN):

Bình luận không có hành vi gấy rối

Ví dụ:

* Mỗi lần mệt mỏi đều mở ra nghe cảm thấy thư giãn và có thêm động lực cảm ơn anh wowy vì bài rap nịnh
* Bài của anh wowy rất hay brother lyrics tuyệt vời sáng tạo
* Theo cách nhìn của mình thì chi hát ổn mà có gì đáng chê

1. Bình luận người dùng phản cảm (OFFENSIVE):

Bình luận có nội dung quấy rối, thậm chí cả những từ ngữ thô tục nhưng không tấn công bất kỳ đối tượng cụ thể nào.

Ví dụ:

* bài hát thấm thía sự trưởng thành ko dành cho trẻ trâu
* Tung hô rap cho cố nào có sức mạnh tinh thần có chiều sâu nhân văn rap nghe n ứng card rap như { i }
* Nghe cái tên đã đại hãm rồi

1. Bình luận người dùng tiêu cực (HATE):

Bình luận có tính chất quấy rối, nội dung lạm dụng và nhắm trực tiếp vào một cá nhân hoặc một nhóm người dựa trên đặc điểm, tôn giáo và quốc tịch.

Ví dụ:

* thằng này giống gay quá
* nờ ô nô đổi tên thành ngờ u ng u

\* Cách gán nhãn dữ liệu:

1. Quy ước:

Theo như bảng định nghĩa ở trên, nhóm chúng em đưa ra bảng quy ước như sau:

|  |  |
| --- | --- |
| Tính chất của bình luận | Nhãn |
| Sạch | CLEAN |
| Phản cảm | OFFENSIVE |
| Tiêu cực | HATE |

1. Ví dụ:

|  |  |
| --- | --- |
| Bình luận | Nhãn |
| * Bài của anh wowy rất hay brother lyrics tuyệt vời sáng tạo | CLEAN |
| * Mỗi lần mệt mỏi đều mở ra nghe cảm thấy thư giãn và có thêm động lực cảm ơn anh wowy vì bài rap nịnh | CLEAN |
| * Tung hô rap cho cố nào có sức mạnh tinh thần có chiều sâu nhân văn rap nghe n ứng card rap như { i } | OFFENSIVE |
| * Nghe cái tên đã đại hãm rồi | OFFENSIVE |
| * thằng này giống gay quá | HATE |
| * nờ ô nô đổi tên thành ngờ u ng u | HATE |

* 1. **Đánh giá độ đồng thuận của dataset**

Dựa trên kết quả sau khi bọn em đã gán nhán trên 500 dữ liệu thì có 485 dữ liệu được gán nhãn giống nhau chiếm 97% trên tổng số dữ liệu. Qua kết quả trên cho thấy sự đồng thuận về dataset cao. Điều này cho thấy mức độ tin cậy và tính rõ ràng của dữ liệu cao. Cho thấy đây là một dữ liệu đáng tin cậy cho quá trình nghiên cứu