CÔNG TY TNHH LẬP TRÌNH VÀ CÔNG NGHỆ CYBERSOFT

Machine Learning

Báo cáo

Phát hiện bình luận tiêu cực

Giáo viên: Nguyễn Việt An

Học viên: Nguyễn Lê Khải Trọng

HO CHI MINH CITY, JANUARY 2026

Contents

1 Code & Video

1.1 Link code github

https://github.com/utbin183/Toxic-Detection

1.2 Link video

2 Giới thiệu

2.1 Giới thiệu vấn đề

Hiện nay, mạng xã hội ngày càng phổ biến, kéo theo số lượng lớn bình luận mang tính tiêu cực, xúc phạm hoặc đe dọa. Tuy nhiên việc kiểm duyệt thủ công tốn nhiều thời gian và không hiệu quả với quy mô lớn. Do đó, việc xây dựng hệ thống phát hiện bình luận độc hại là cần thiết.

2.2 Mục tiêu dự án

Là xây dựng hệ thống phát hiện bình luận tiêu cực dựa trên 2 mô hình học máy là Logistic Regression và LSTM để có thể so sánh hiệu quả giữa 2 mô hình và triển khai mô hình trên ứng dụng streamlit để người dùng dễ tương tác.

2.3 Phạm vi bài toán

Bài toán được xây dựng dưới dạng phân loại nhị phân:

- 0: Không độc hại
- 1: Độc hại

3 Khái niệm và Giải thích mô hình

3.1 Nhu cầu giải thích trong bài toán phát hiện bình luận độc hại

Trong các hệ thống phát hiện bình luận tiêu cực, ngoài độ chính xác, khả năng giải thích lý do vì sao một bình luận bị phân loại là *toxic* là một yếu tố quan trọng. Việc giải

thích giúp người dùng hiểu được hành vi của mô hình, đồng thời tăng mức độ tin cậy khi áp dụng trong thực tế.

Đối với bài toán kiểm duyệt nội dung, các quyết định của mô hình có thể ảnh hưởng trực tiếp đến trải nghiệm người dùng. Do đó, việc phân tích các yếu tố tác động đến kết quả dự đoán là cần thiết, đặc biệt trong bối cảnh dữ liệu có thể mang tính nhạy cảm.

3.2 Từ khoá kích hoạt (Trigger Words)

Từ khoá kích hoạt là các từ hoặc cụm từ trong bình luận có mức độ ảnh hưởng lớn đến xác suất mô hình dự đoán một bình luận là độc hại. Thông thường, đây là các từ mang tính xúc phạm, đe doạ hoặc công kích cá nhân.

Ví dụ về các từ khoá thường xuất hiện trong các bình luận toxic:

- *idiot, stupid, loser*
- *shut up, kill, hate*

Việc xác định các từ khoá kích hoạt giúp phân tích xem mô hình đang tập trung vào những yếu tố nào trong quá trình ra quyết định, đồng thời hỗ trợ việc phát hiện các trường hợp mô hình học sai hoặc bị thiên lệch.

3.3 Giải thích mô hình Baseline (TF-IDF + Logistic Regression)

Đối với mô hình Logistic Regression kết hợp với TF-IDF, việc giải thích tương đối trực quan. Mỗi từ trong văn bản sau khi vector hoá sẽ tương ứng với một trọng số trong mô hình.

Các từ có trọng số dương lớn cho thấy chúng có xu hướng làm tăng khả năng bình luận bị phân loại là toxic. Ngược lại, các từ có trọng số âm thường liên quan đến các bình luận không độc hại.

Thông qua việc trích xuất các từ có trọng số cao nhất, có thể nhận thấy mô hình Baseline thường tập trung vào các từ xúc phạm rõ ràng. Tuy nhiên, do không xét đến ngữ cảnh, mô hình có thể đưa ra dự đoán sai trong các trường hợp mỉa mai hoặc sử dụng từ ngữ theo nghĩa tích cực.

3.4 Giải thích mô hình Deep Learning (LSTM) ở mức cơ bản

Mô hình LSTM thuộc nhóm mô hình "hộp đen", do đó việc giải thích trực tiếp các quyết định của mô hình trở nên khó khăn hơn so với các mô hình truyền thống. Trong

phạm vi dự án, việc giải thích được thực hiện ở mức cơ bản, tập trung vào các phương pháp phổ biến.

3.4.1 Attention Weights

Cơ chế Attention cho phép mô hình gán mức độ quan trọng khác nhau cho từng từ trong câu. Các từ có giá trị attention cao thường là những từ có ảnh hưởng lớn đến kết quả dự đoán.

Trong dự án này, Attention được xem xét chủ yếu ở mức khái niệm, nhằm minh hoạ cách mô hình Deep Learning có thể tập trung vào các từ khoá quan trọng, do giới hạn về thời gian và phạm vi triển khai.

3.4.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME là một phương pháp giải thích độc lập với mô hình, hoạt động bằng cách tạo ra các biến thể của đầu vào và quan sát sự thay đổi trong đầu ra của mô hình.

Thông qua LIME, có thể xác định được các từ trong một bình luận cụ thể có ảnh hưởng tích cực hoặc tiêu cực đến quyết định dự đoán. Phương pháp này đặc biệt hữu ích khi phân tích từng trường hợp riêng lẻ.

3.4.3 SHAP (SHapley Additive exPlanations)

SHAP dựa trên lý thuyết trò chơi để ước lượng mức đóng góp của từng đặc trưng vào kết quả dự đoán cuối cùng. Mỗi từ trong câu sẽ có một giá trị SHAP thể hiện mức độ ảnh hưởng của từ đó đến nhãn toxic.

Trong dự án này, SHAP được đề cập như một hướng tiếp cận tiềm năng để mở rộng khả năng giải thích mô hình trong các nghiên cứu tiếp theo.

3.5 Nhận xét và hạn chế

Các phương pháp giải thích mô hình giúp cung cấp cái nhìn tổng quan về cách mô hình đưa ra quyết định. Tuy nhiên, vẫn tồn tại một số hạn chế:

- Khó giải thích đầy đủ các trường hợp mỉa mai hoặc ẩn ý
- Một số từ trung tính có thể bị đánh giá sai do thiếu ngữ cảnh

Việc tích hợp các phương pháp giải thích nâng cao hơn sẽ là một hướng phát triển trong tương lai.

Tiêu chí	Logistic Regression	LSTM
Tốc độ	Nhanh	Chậm
Hiểu ngữ cảnh	Kém	Tốt
Độ chính xác	Trung bình	Cao hơn

Table 4.1: Bảng so sánh 2 models

3.6 Liên hệ với vấn đề đạo đức

Khả năng giải thích mô hình đóng vai trò quan trọng trong việc đảm bảo tính minh bạch và công bằng. Việc hiểu được lý do mô hình đưa ra quyết định giúp giảm nguy cơ thiên lệch, đồng thời hỗ trợ con người trong quá trình kiểm duyệt nội dung, thay vì thay thế hoàn toàn yếu tố con người.

4 Đánh giá và kết quả

4.1 Thước gia đánh giá

- F1-score
- ROC-AUC
- PR-AUC
- Confusion Matrix

4.2 Kết quả thực nghiệm

- Baseline đạt F1-score ở mức khá, tốc độ nhanh.
- LSTM cho kết quả tốt hơn về F1 và ROC-AUC
- Deep Learning mô hình hiểu ngữ cảnh tốt hơn trong các câu dài

4.3 So sánh 2 mô hình

5 Phân tích lỗi

5.1 Phân tích lỗi

- False Positive là những câu có từ ngữ mạnh và không mang tính xúc phạm. Ví dụ: "This game is sick"
- False Negative là câu mỉa mai, toxic. Ví dụ: "You are bad"

6 Kết luận

Dự án đã xây dựng thành công hệ thống phát hiện bình luận độc hại và kết quả cho thấy deep learning hiệu quả hơn baseline.