PHÁT HIỆN BÌNH LUẬN TIÊU CỰC TRÊN MẠNG XÃ HỘI Ở VIỆT NAM

Trần Xuân Minh[1,1]

[1] 21520352@gm.uit.edu.vn

[2]Trường Đại học Công Nghệ Thông Tin ĐHQG TP.HCM

Tóm tắt

Nghiên cứu này tập trung vào việc phát triển một hệ thống có khả năng phát hiện bình luận tiêu cực:

- Tạo ra bộ dữ liệu mới Vi-COHSC cho tiếng Việt.
- Nghiên cứu các phương pháp tiếp cận tiên tiến Bi-LSTM, RoBERTa, PhoBERT-CNN.
- Xây dựng một ứng dụng Web phát hiện và loại bỏ các bình luận tiêu cực.

Lý do chọn đề tài

- Sự tự do ngôn luận trên mạng xã hội và thiếu nhân lực trong nhiệm vụ kiểm duyệt nội dung dẫn đến nhiều nội dung độc hại vi phạm tiêu chuẩn cộng động đang gia tăng nhanh chóng ảnh hưởng tiêu cực đến hành vi con người cũng như ảnh hưởng trực tiếp tới xã hội.
- Góp phần hoàn thiện nhiệm vụ phát hiện bình luận tiêu cực mang tính thù hận và xúc phạm cho ngôn ngữ Tiếng Việt.

Giới thiệu

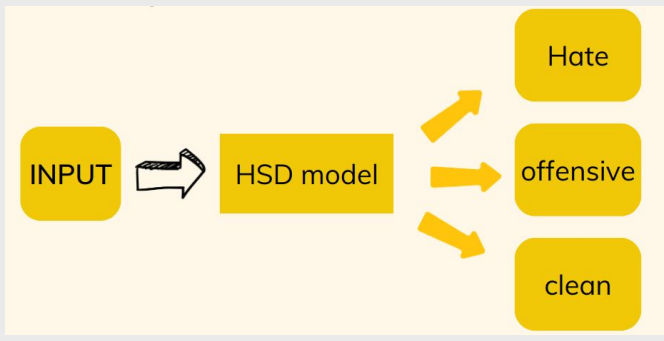
Vấn đề

Input

Những bình luận tiếng Việt trên các trang mạng xã hội

Output

Một trong ba nhãn khác nhau được các bộ phân loại dự đoán.



Nội dung và Phương pháp

1. Tạo bộ dữ liệu mới

- Thu thập bình luận của người dùng về các vấn đề xã hội và chính trị từ các trang Facebook và YouTube khác nhau ở Việt Nam
- xây dựng quy trình tiền xử lý dữ liệu để cải thiện chất lượng của bộ dữ liệu này, nhằm trích xuất các đặc trưng có giá trị như loại bỏ những khoảng trống không cần thiết, xóa liên kết
- Đánh giá tập dữ liệu dựa trên tính toán sự đồng thuận giữa các nhà chú thích bằng trị số Kappa của Cohen (κ)

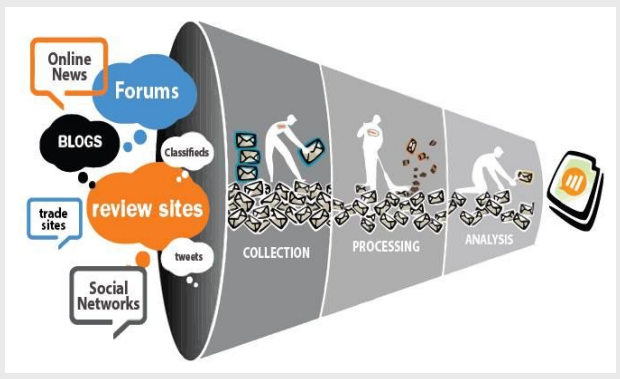


Figure 1. Hình minh họa quá trình thu nhập và xây dựng bộ dữ liệu

2. Huấn luyện các mô hình

- Huấn luyện các mô hình Bidirectional Long Short-Term Memory (Bi-LSTM), Robustly Optimized BERT Approach (RoBERTa), và PhoBERT-CNN.
- Sử dụng độ đo Accuracy và F1-score để so sánh và đánh giá hiệu suất của các phương pháp được đề xuất.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 2. Hình minh họa so sánh và đánh giá các phương pháp được đề xuất

3. Xây dựng ứng dụng Web

- Áp dụng các kỹ thuật Crawl Data từ trang web, lưu trữ thông tin về ID của bài đăng, hoặc URL của bài viết.
- Dự đoán nhãn cho các bình luận và báo cáo lại cho quản trị viên nếu có bình luận nào được gắn nhãn là HATE hoặc OFFENSIVE.

Figure 3. Xây dựng ứng dụng Web phát hiện và loại bỏ bình luận tiêu cực

Trần Xuân Minh– Trường Đại học Công Nghệ Thông Tin ĐHQG TP.HCM
TEL : 0344542073 Email : 21520352@gm.uit.edu.vn