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

Trần Xuân Minh - 21520352

Tóm tắt

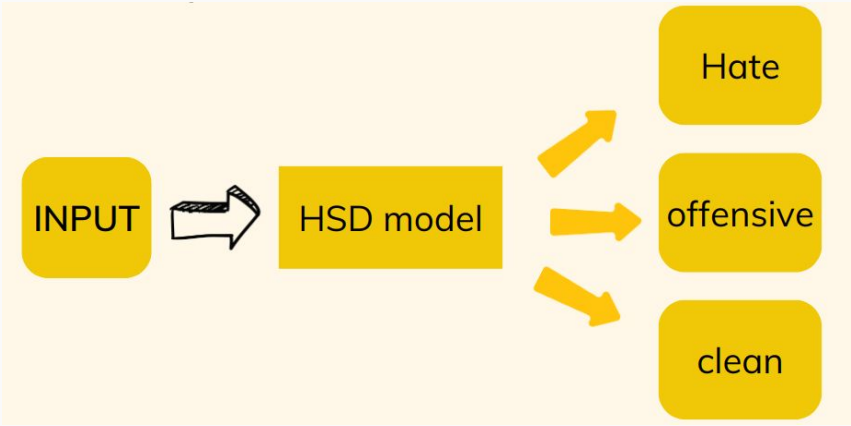
- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/TxmMinh/Report-CS519.O11
- Link YouTube video: https://youtu.be/pykwxqYnIdA

Trần Xuân Minh
21520352

Giới thiệu

- Vấn đề: nhiều nội dung độc hại vi phạm tiêu chuẩn cộng đồng đang gia tăng nhanh chóng trên mạng xã hội ở Việt Nam.

- 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.



Mục tiêu

- Tạo ra bộ dữ liệu mới Vi-COHSC cho tiếng Việt, với các chú thích gán nhẫn nghiêm ngặt và quy trình đánh giá tập dữ liệu để đo lượng sự đồng thuận giữa các nhà chú thích.
- Nghiên cứu các phương pháp học sâu Bidirectional Long Short-Term Memory, phương pháp học chuyển tiếp Robustly optimized BERT approach và phương pháp học kết hợp PhoBERT-CNN.
- Xây dựng ứng dụng Web tích hợp vào các trang mạng xã hội để phát hiện và loại bỏ những bình luận tiêu cực mang tính căm thù và xúc phạm.

Nội dung và Phương pháp

- 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. Sau đó tiến hành xử lý bộ dữ liệu.
- Huấn luyện các mô hình Bidirectional Long Short-Term Memory (Bi-LSTM), Robustly Optimized BERT Approach (RoBERTa), và PhoBERT-CNN.
- 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 bình luận được dự đoán độc hại.

Kết quả dự kiến

- Cung cấp một bộ dữ liệu Vi-COHSC chất lượng cao cho cộng đồng nghiên cứu tiếng Việt.
- Các phương pháp Bidirectional Long Short-Term Memory (Bi-LSTM), Robustly Optimized BERT Approach (RoBERTa), và phương pháp kết hợp PhoBERT-CNN đạt kết quả thực nghiệm tốt.
- Ứng dụng Web đạt hiệu suất tốt trong việc tự động phát hiện và loại bỏ bình luận tiêu cực mang tính căm thù và xúc phạm.

Tài liệu tham khảo

- [1] Khanh Q. Tran, An T. Nguyen, Phu Gia Hoang, Canh Duc Luu, Trong-Hop Do, Kiet Van Nguyen. Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data. arXiv preprint arXiv:2206.00524, 2022.
- [2] Luan Thanh Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen. Constructive and Toxic Speech Detection for Open-domain Social Media Comments in Vietnamese. arXiv preprint arXiv:2103.10069, 2021.
- [3 Son T. Luu, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen. A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts. arXiv preprint arXiv:2103.11528, 2021.
- [4] Hang Thi-Thuy Do, Huy Duc Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, Anh Gia-Tuan Nguyen. Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model. arXiv preprint arXiv:1911.03648, 2019.