THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: *https://youtu.be/pykwxqYnIdA*
- Link slides: *https://github.com/TxmMinh/Report-CS519.O11*

• Họ và Tên: Trần Xuân Minh
• MSSV: 21520352
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 9/10
	• Số buổi vắng: 1
	• Số câu hỏi QT cá nhân: 15/15
	• Link Github: https://github.com/mynameuit/CS519.O11/
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng cho đề tài
	  ○ Viết đề cương, slides, poster
	  ○ Làm video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DETECTION OF NEGATIVE COMMENTS ON SOCIAL MEDIA IN VIETNAM

TÓM TẮT

Nghiên cứu này tập trung vào việc phát triển một hệ thống thông minh có khả năng giải quyết những thiếu sót lớn về thiếu nhân lực trong việc kiểm duyệt nội dung có trình độ cao về Tiếng Việt, mô hình hiệu suất khiêm tốn và thiếu ứng dụng thực tế. Đầu tiên, xây dựng bộ dữ liệu mới Vi-COHSC cho tiếng Việt được thu thập từ mạng xã hội Việt Nam với các nguyên tắc chú thích cho các bình luận và quy trình đánh giá bộ dữ liệu nghiêm ngặt. Thứ hai, nghiên cứu các phương pháp tiếp cận tiên tiến: phương pháp học sâu Bidirectional Long Short-Term Memory (Bi-LSTM), phương pháp học chuyển tiếp Robustly optimized BERT approach (RoBERTa) và phương pháp học kết hợp PhoBERT-CNN. So sánh và đánh giá hiệu suất của các mô hình đề xuất trên bộ dữ liệu Vi-COHSC. Cuối cùng, xây dựng một ứng dụng Web phát hiện và loại bỏ các bình luận tiêu cực để chứng minh tính thực tiễn của hệ thống được đề xuất.

GIỚI THIỆU

Những năm gần đây, mạng xã hội ngày càng phát triển và dần trở nên phổ biến với mọi người trên thế giới như Facebook, YouTube, Instagram. Ở Việt Nam, số lượng người dùng mạng xã hội vẫn đang tiếp tục gia tăng mà không có dấu hiệu chững lại. Tuy nhiên, với sự tự do ngôn luận, bình 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 đồng mang tính chất thù hận và xúc phạm đang gia tăng nhanh chóng ảnh hưởng tiêu cực đến hành vi con người cũng như ảnh hưởng trực tiếp tới xã hội.

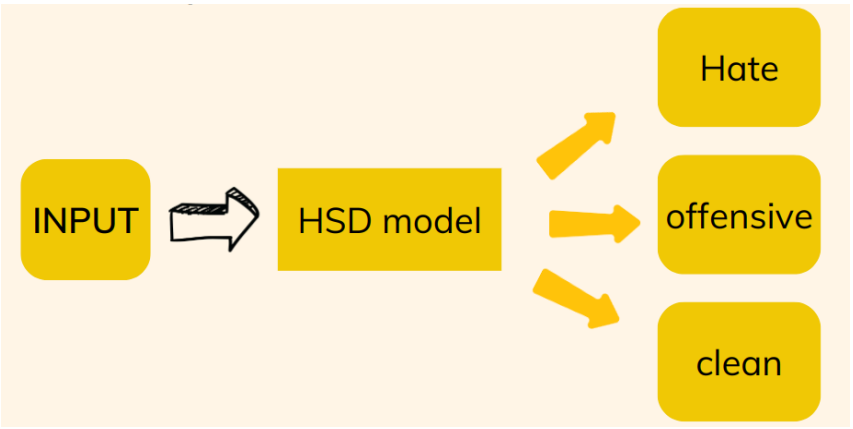
Các tập đoàn lớn trên thế giới như Facebook, YouTube sử dụng các phương pháp tiên

tiến áp dụng trên đa ngôn ngữ nhằm tạo ra các hệ thống mạnh mẽ giúp phân loại nội dung tiêu cực. Tuy nhiên, các hệ thống này khó có thể nhận ra nội dung thiếu bối cảnh, sự khác biệt về văn hóa ở mỗi Quốc gia và chậm theo kịp sự phát triển liên tục của nội dung độc hại.

Để góp phần hoàn thiện nhiệm vụ phát hiện bình luận tiêu cực mang tính thù hận và xúc phạm cho ngôn ngữ Tiếng Việt. Đó là động lực để chúng tôi tạo ra một bộ dữ liệu mới mang tên Vi-COHSC cho tiếng Việt với các nguyên tắc chú thích cho các bình luận và quy trình đánh giá nghiêm ngặt để đo lường sự đồng thuận giữa các nhà chú thích. Trong đề tài này, chúng tôi 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 và áp dụng chúng trên bộ dữ liệu của chúng tôi nhằm phát hiện xem các bình luận trên mạng xã hội là THÙ GHÉT (HATE), CÔNG KÍCH (OFFENSIVE) hay SẠCH (CLEAN). Cụ thể:

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.



• Căm thù (HATE) chứa ngôn ngữ lăng mạ, thường có mục đích xúc phạm các cá nhân hoặc nhóm và có thể bao gồm lời nói căm thù, ngôn từ xúc phạm. Ví dụ: Cứ phải chửi cho mới chịu ngậm mồm, lũ cẩu đói.

• Xúc phạm nhưng không căm thù (OFFENSIVE) là các bình luận có nội dung quấy rối, thậm chí có từ ngữ tục tĩu nhưng không công kích bất kỳ đối tượng cụ thể nào. Ví dụ: Đồ khùng.

• Không phải xúc phạm hay căm ghét (CLEAN) là một bình luận bình thường và

không chứa ngôn ngữ xúc phạm hoặc lời nói căm thù. Ví dụ: Hôm nay trời đẹp!

MỤC TIÊU

+ Các nghiên cứu hiện tại về phát hiện bình luận tiêu cực không tập trung vào việc phân tích khía cạnh cảm xúc của ngôn ngữ tiếng Việt vậy nên chúng tôi tạo ra một bộ dữ liệu mới mang tên 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 mô hình phát hiện bình luận tiêu cực dựa trên các phương pháp tiếp cận tiên tiến: phương pháp học sâu Bidirectional Long Short-Term Memory (Bi-LSTM), phương pháp học chuyển tiếp Robustly optimized BERT approach (RoBERTa) và phương pháp học kết hợp PhoBERT-CNN, là sự kết hợp giữa mô hình PhoBERT và mô hình Text-CNN. Sau đó, so sánh hiệu suất của các mô hình này bằng các độ đo đánh giá Accuracy và F1-score.

+ Xây dựng ứng dụng áp dụng mô hình được nghiên cứu, sau đó 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

a. NỘI DUNG:

+ Tự xây dựng bộ dữ liệu mới Vi-COHSC cho tiếng Việt, với các chú thích gán nhẫn nghiêm ngặt để mô hình có thể huấn luyện tốt từ bộ dữ liệu.

+ Nghiên cứu các phương pháp tiếp cận tiên tiến: Bidirectional Long Short-Term Memory (Bi-LSTM), Robustly Optimized BERT Approach (RoBERTa), và phương pháp kết hợp PhoBERT-CNN.

+ Xây dựng ứng dụng web phát hiện và xóa bỏ các bình luận tiêu cực.

b. PHƯƠNG PHÁP:

+ Thu thập nhận xét, 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. Chọn lọc các trang Facebook và kênh YouTube có tỷ lệ tương tác cao. Sau khi thu thập dữ liệu, 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 bởi việc có link website trong bình luận không ảnh hưởng đến cảm xúc của bình luận, loại bỏ các ký tự dư thừa mà người dùng cố ý tạo ra,... Bộ dữ liệu Vi-COHSC chứa ba nhãn: HATE, OFFENSIVE và CLEAN. Mỗi nhà chú thích gán một nhãn cho mỗi bình luận trong bộ dữ liệu. Nếu có bất kỳ nhãn nào khác nhau giữa hai nhà chú thích, bổ sung nhà chú thích thứ ba chú thích những nhãn đó. Nhà chú thích thứ tư chú thích nếu tất cả ba nhà chú thích đều không đồng ý. Nhãn cuối cùng được xác định bằng cách bầu chọn dựa trên đa số. Bằng cách này, mỗi bình luận được chú thích bằng một nhãn và tính khách quan cho mỗi bình luận. Sau đó đánh giá tập dữ liệu dựa trên tính toán sự đồng thuận giữa các nhà chú thích bằng trị số Kappa của Cohen (κ). Nếu sự đồng thuận không đủ tốt, huấn luyện lại những người chú thích và cập nhật lại hướng dẫn chú thích nếu cần thiết.

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

+ Xây dựng một ứng dụng Web tự động thu thập các bình luận trên các trang mạng xã hội và gán nhãn các bình luận tiêu cực độc hại. Sau đó các bình luận được gắn nhãn độc hại sẽ được báo cáo lại cho quản trị viên của trang mạng xã hội như nhóm Facebook hoặc kênh YouTube đó để xóa các bình luận và chặn người dùng nào đã bình luận. Để thực hiện được điều đó, chúng tôi nghiên cứu các kỹ thuật Crawl Data từ một trang web như sử dụng các thư viện hoặc công cụ crawl như BeautifulSoup hoặc Scrapy để thu thập dữ liệu từ các trang mạng xã hội. Ngoài ra, áp dụng kỹ thuật crawler dữ liệu website sử dụng kỹ thuật phân tích cú pháp XML bằng PHP. Đối với Facebook, nghiên cứu API Facebook Graph để lấy dữ liệu bình luận từ các bài đăng hoặc trang, nhóm. Để xác định bài viết hoặc người nào đã bình luận, cần lưu trữ thông tin về ID của bài đăng, hoặc URL của bài viết. Sau đó, tiến hành 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.

KẾT QUẢ MONG ĐỢI

+ Cung cấp một bộ dữ liệu Vi-COHSC chất lượng cao cho cộng đồng nghiên cứu tiếng Việt.

+ Báo cáo các kỹ thuật, kết quả thực nghiệm, so sánh và đánh giá của các phương pháp Bidirectional Long Short-Term Memory (Bi-LSTM), Robustly Optimized BERT Approach (RoBERTa), và phương pháp kết hợp PhoBERT-CNN.

+ Ứng dụng Web đạt hiệu suất tốt trong việc tự động phát hiện bình luận tiêu cực mang tính căm thù và xúc phạm. Báo cáo các bình luận này cho quản trị viên của trang mạng xã hội liên quan để ngăn chặn người dùng và xóa bỏ bình luận.

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.