Vietnamese harmful speech detection and Youtube comments classfication

Đoàn Minh Tuấn
Trường Đại học Công nghệ thông tin
MSSV: 22521600
Email: 22521600@gm.uit.edu.vn

***Tóm tắt nội dung*—Sự hiện diện phổ biến của mạng xã hội trực tuyến mang lại nhiều hệ quả tích cực và tiêu cực cho xã hội. Bên cạnh những lợi ích, mạng xã hội cũng có thể gây ra những vấn đề lớn do nội dung thù hận và xúc phạm. Việc phát hiện và loại bỏ những nội dung độc hại này bằng cách sử dụng học máy là một chủ đề nghiên cứu chính trong mạng xã hội. Hai trong số những thách thức của chủ đề này là khối lượng dữ liệu mạng xã hội rất lớn và dữ liệu này cần được xử lý trong thời gian thực. Trong bài báo này, chúng tôi đặt mục tiêu phát triển hệ thống phát hiện lời nói độc hại trong các bình luận trên YouTube tiếng Việt bằng cách sử dụng công nghệ học máy và công nghệ dữ liệu lớn. Dữ liệu phát trực tuyến từ YouTube được xử lý trong thời gian thực bằng cách sử dụng Spark và công nghệ học máy. Cuối cùng, thư viện Flask sẽ được sử dụng để hiển thị kết quả.**

***Index Terms*—máy học, dữ liệu lớn, lời nói độc hại, mạng xã hội trực tuyến.**

I. Giới thiệu

Mạng xã hội là một trong những công nghệ hoàn toàn thay đổi xã hội. Bên cạnh nhiều lợi ích lớn, mạng xã hội cũng gây ra nhiều vấn đề, trong đó lạm dụng trực tuyến là một trong những vấn đề được quan tâm nhất. Loại lạm dụng trực tuyến phổ biến nhất là việc đưa ra các bình luận thù hận và xúc phạm trên mạng xã hội. Mặc dù người dùng có thể báo cáo những bình luận này, giải pháp cuối cùng là xây dựng một hệ thống dựa trên trí tuệ nhân tạo để tự động phát hiện và lọc bỏ những nội dung độc hại này khỏi mạng xã hội.

Trong những năm gần đây, phát hiện lời nói thù hận là một chủ đề nghiên cứu tích cực đã thu hút nhiều sự chú ý từ giới học thuật.

Tính thực tiễn của hệ thống phát hiện lời nói thù hận là mối quan tâm chính trong nghiên cứu này. Do đó, thay vì cố gắng điều chỉnh các mô hình học máy để cải thiện hiệu suất phát hiện lời nói thù hận, nghiên cứu này tập trung vào việc trình bày một triển khai của hệ thống phát hiện lời nói thù hận cho dữ liệu truyền phát từ mạng xã hội thực sự khả thi. Điều này có nghĩa là hệ thống có khả năng xử lý một lượng lớn dữ liệu truyền phát từ các bình luận trên mạng xã hội và tạo ra kết quả trong thời gian thực. Cụ thể hơn, các bình luận từ mạng xã hội được thu thập và lưu trong một file csv sau đó được đưa vào một mô hình phát hiện lời nói thù hận đã được huấn luyện, tích hợp vào Structured Streaming bên trong Spark, một khung công tác mạnh mẽ cho xử lý dữ liệu lớn. Việc sử dụng Spark Streaming cho phép xử lý một lượng lớn bình luận từ mạng xã hội và xuất kết quả trong thời gian thực. Kết quả sau đó được hiển thị thông qua các biểu dựa trên ứng dụng web.

Tôi sẽ cung cấp Bộ dữ liệu trong Phần II và Phương pháp và hướng tiếp cận trong Phần III Sau đó, trong phần IV, tôi sẽ đánh giá kết quả. Cuối cùng, trong phần V, tôi sẽ đưa ra kết luận và đề xuất hướng phát triển trong tương lai.

II. Bộ dữ liệu

A. Giới thiệu bài toán

Trong phần này, chúng tôi tóm tắt Nhiệm vụ Phát hiện lời nói độc hại trong tiếng Việt. Nhiệm vụ này nhằm phát hiện xem một bình luận trên mạng xã hội là độc hại hay không hay sạch sẽ. Bài toán được mô tả như sau.

Đầu vào: Một comment là tiếng Việt trên Youtube.

Đầu ra: Dự đoán một bình luận là **CLEAN** hay **HARMFUL**.

Trong đó:

- **Harmful speech**: chứa những bình luận là xúc phạm hay gây thù hận.
- **Clean speech**: cuộc trò chuyện, thể hiện cảm xúc một cách bình thường. Nó không chứa ngôn ngữ xúc phạm hoặc lời nói thù hận.

B. Tổng quan

Tôi sử dụng bộ dữ liệu được lấy từ Vietnamese Hate Speech Detection(Lưu et al., 2021) [1]. Tập dữ liệu ViHSD bao gồm 33.400 bình luận trên mạng xã hội. Tập dữ liệu này được chia thành các tập train, dev và test. Mỗi dòng dữ liệu sẽ được gán một trong ba nhãn: CLEAN (sạch), OFFENSIVE (xúc phạm) hoặc HATE (thù hận). Có sự chênh lệch lớn về số lượng bình luận được gán nhãn CLEAN(bình thường) so với bình luận được gán nhãn OFFENSIVE(xúc phạm) và HATE(gây thù). Bên cạnh đó, chúng tôi nhận thấy rằng các bình luận OFFENSIVE và HATE đều có đặc điểm chung là những bình luận độc hại ảnh hưởng tới người dùng nên tôi quyết định sẽ gộp 2 nhãn OFFENSIVE và HATE thành 1 nhãn chung là HARMFUL. Bảng 1 trình bày tổng quan về tập dữ liệu này.

C. Tiền xử lí dữ liệu

Bước tiền xử lý dữ liệu là một bước quan trọng trong hầu hết các dự án Máy học (Machine Learning). Do đó, để làm

Bảng I
TỔNG QUAN BỘ DỮ LIỆU

Comments	Nhãn
Em được làm fan cứng luôn rồi nè ♡ reaction quá hay quá cute coi mấy giờ này quá hợp lí =]]]	CLEAN
Quá ngu lồn đi =)))	OFFENSIVE
mài có óc để suy nghĩ ko add? cái đồng băng đó mài tính đào hồ bnhiu cho đủ hả thằng ngu?	HATE

sạch bộ dữ liệu đã cho một cách tốt nhất, tôi đề xuất một quy trình tiền xử lý dữ liệu hai giai đoạn.

Giai đoạn 1: Tôi sẽ xóa các biểu tượng emoji, đường link, urls, kí tự đặc biệt, hashtag, chữ lặp lại và khoảng trắng dư thừa. Ngoài ra, tôi sẽ chuẩn hóa lại các từ viết tắt. Ví dụ: adidaphat -> a đi đà phật, ak -> à, ae -> anh em.

Giai đoạn 2: Tôi sẽ sử dụng thư viện Pyvi để phân tách comment thành các từ hoặc cụm từ. Sau đó tôi cũng loại bỏ stopwords khỏi các bình luận vì nó là những từ hoặc cụm không mang nhiều ý nghĩa. Trong các thí nghiệm của chúng tôi, chúng tôi sử dụng bộ dữ liệu Vietnamese stopword dictionary [2] để loại bỏ các stopwords trong câu.

III. PHƯƠNG PHÁP VÀ HƯỚNG TIẾP CẬN

A. Mô hình máy học

Trong đề tài, tôi sẽ sử 2 mô hình máy học để dự đoán bình luận trên Youtube.

1) **Logistic Regression**: là một phương pháp thống kê được sử dụng để dự đoán mối quan hệ giữa một biến phụ thuộc nhị phân (có hai giá trị, chẳng hạn như 0 và 1) và một hoặc nhiều biến độc lập (có thể là rời rạc hoặc liên tục).
2) **Decision Tree**: là một mô hình học máy dùng để ra quyết định thông qua việc phân chia dữ liệu dựa trên các thuộc tính. Nó được biểu diễn dưới dạng một cấu trúc cây, với nút gốc đại diện cho dữ liệu ban đầu, các nút nội thể hiện các thuộc tính, và các nút lá cho kết quả cuối cùng. Quy trình phân tách dựa trên các điều kiện cho phép cây quyết định phân loại dữ liệu hoặc dự đoán giá trị.

B. Tham số chính của mô hình máy học

Đầu tiên, tôi sẽ sử dụng kỹ thuật TF – IDF (Term Frequency-Inverse Document Frequency) kết hợp với tham số N-gram = 2 để chuyển đổi dữ liệu văn bản thành các vector.

- Logistic Regression: max_Iter = 20, regPram = 0.3
- Decision Tree: maxDepth = 17, minInstancesPerNode = 3

Trong đó:

- max_Iter: số lần lặp tối đa mà thuật toán sẽ thực hiện để tìm ra các tham số tối ưu cho mô hình.
- regPram(hay là C): tham số điều chỉnh độ mạnh của quy định hóa (regularization) trong mô hình.
- maxDepth: độ sâu tối đa của cây quyết định.
- minInstancesPerNode: số lượng mẫu tối thiểu mà một nút trong cây quyết định phải có trước khi nó có thể được phân chia tiếp.

C. Mô hình hệ thống

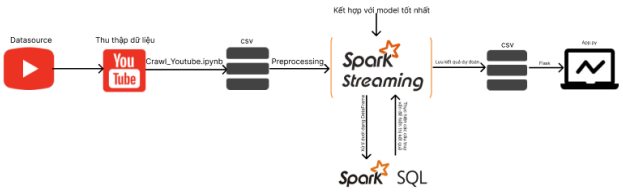
- Apache Spark: là một framework mã nguồn mở được sử dụng để xử lý và phân tích dữ liệu lớn một cách nhanh chóng và hiệu quả. Spark được thiết kế để xử lý dữ liệu theo phương thức phân tán trên nhiều máy tính, cho phép xử lý dữ liệu lớn (Big Data) với hiệu suất cao nhờ vào khả năng lưu trữ dữ liệu trong bộ nhớ.
- Flask: là một framework web nhỏ gọn và linh hoạt cho Python. Nó được thiết kế để giúp xây dựng các ứng dụng web nhanh chóng và dễ dàng.

D. Hướng tiếp cận và xây dựng hệ thống

1) Thu thập bình luận từ Youtube Data API
   - Trong phần này, tôi đã xác thực và kết nối với Youtube Data API bằng cách sử dụng thông tin xác thực của nhà phát triển(developer credentials). Đầu tiên, tôi cần thiết lập thông tin thiết yếu để đăng nhập và sử dụng Youtube Data API, chẳng hạn như DEVELOPER_KEY, YOUTUBE_API_SERVICE_NAME, YOUTUBE_API_VERSION.
   - Sau đó tôi sẽ lấy id của 1 video bất kì trên youtube để thu thập comment. Trong đề tài, tôi sẽ lấy id của video có tiêu đề là: **MAN UNITED - SOUTHAMPTON | NGƯỜI HÙNG AMAD DIALLO, VÕ ÒA 10 PHÚT CUỐI CÙNG | NGOẠI HẠNG ANH 24/25**. Phần id của video sẽ nằm sau v= của đường link video đấy. Ví dụ, đường link video trong đề tài là: https://www.youtube.com/watch?v=Iex2yOj99Q0 thì phần id sẽ là: Iex2yOj99Q0.
2) Xử lí dữ liệu đã thu thập được
   Đầu tiên, tôi sẽ tiền xử lí dữ liệu đã thu thập dựa vào quá trình tiền xử lí đã nói ở trên mục C của phần II. Đưa dữ liệu đã được tiền xử lí vào Spark Streaming sau đó áp dụng mô hình tốt nhất đã được lưu để dự đoán kết quả là CLEAN hay HARMFUL. Sau khi có kết quả dự đoán, sử dụng SparkSQL để hiển thị kết quả dự đoán. Lưu kết quả đã dự đoán được vào một file csv. Cuối cùng, sử dụng thư viện flask tạo nên một trang web để hiển dữ liệu kết quả dự đoán đã được lưu một cách trực quan hơn. Hình 1 sẽ biểu diễn mô hình hệ thống.

IV. KẾT QUẢ THU ĐƯỢC

Phần này sẽ các chỉ số đánh giá được sử dụng trong đề tài này. Accuracy và average macro F1-score là các chỉ số phổ biến và được sử dụng rộng rãi cho các nhiệm vụ phân loại nói chung và xác định các bình luận thù hận và khiếm nhã nói riêng. Tuy nhiên, do các lớp không cân bằng đáng kể trong các tập dữ liệu được cung cấp, average macro F1-score là chỉ số phù hợp nhất cho đề tài. Kết quả là, khi đánh giá hiệu

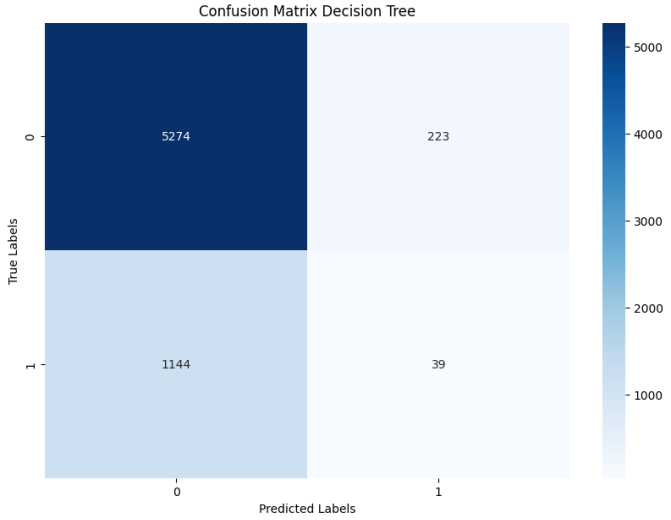


Hình 1. Mô hình hệ thống

suất của mô hình, chúng tôi đã chọn sử dụng average macro F1-score (%) làm chỉ số chính và Accuracy (%) để cung cấp thông tin bổ sung ở Bảng II. Ngoài ra, nếu các độ đo đánh giá của 2 mô hình không có sự chênh lệch lớn, tôi sẽ sử dụng thêm Confussion Matrix ở Hình 2 và Hình 3 để đánh giá thêm hiệu suất của 2 mô hình phân loại bằng cách hiển thị số lượng mẫu được phân loại đúng và sai.

Bảng II
KẾT QUẢ ĐỘ ĐO ĐÁNH GIÁ TRÊN TẬP TEST

Method	F1-score	Accuracy
Decision Tree	0.74	0.8
Logistic Regression	0.74	0.81



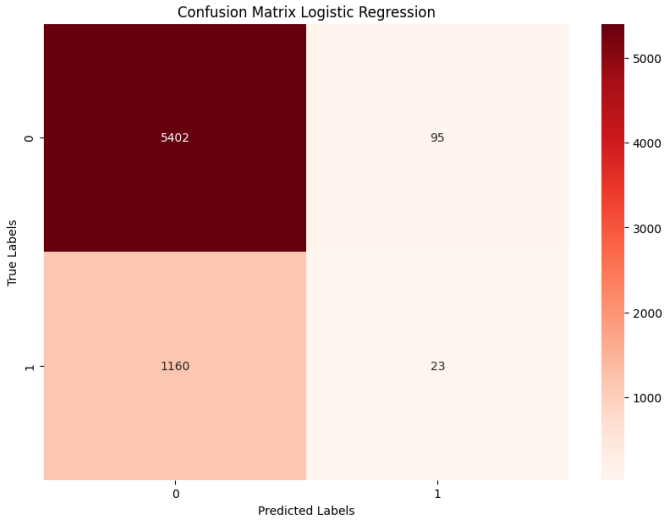
Hình 2. Decision Tree Confussion Matrix

Ở bảng II, ta có thể thấy F1-score và Accuracy ở 2 mô hình không chênh lệch nhiều tuy nhiên ở Confussion Matrix Decision Tree dự đoán các nhãn CLEAN và HARMFUL tốt hơn so với với Logistic Regression nên tôi sẽ chọn mô hình Decision Tree để dự đoán comment trên Youtube.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

A. Kết luận

Đã triển khai được hệ thống phát hiện các comment độc hại trên nền tảng mạng xã hội bằng Apache Spark cũng như sử dụng được các mô hình học máy trong Spark Mllib. Ngoài ra, còn biết thêm được về kỹ thuật xử lí ngôn ngữ tự nhiên (NLP).



Hình 3. Logistic Regression Confussion Matrix

B. Hướng phát triển

Sử dụng thêm các mô hình DeepLearning trong xử lí ngôn ngữ tự nhiên như: PhoBert_CNN,... Ngoài ra có thể nghiên cứu để triển khai đồ án ở chế độ xử lí song song để nâng cao chất lượng tài nguyên cũng như cải thiện hiệu suất tính toán.

TÀI LIỆU

[1] Luu, S.T., Nguyen, K.V., Nguyen, N.L.-T.: A large-scale dataset for hate speech detection on vietnamese social media texts. In: Fujita, H., Selamat, A., Lin, J.C.-W., Ali, M. (eds.) Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices, pp. 415–426. Springer, Cham (2021)
[2] Le, V.-D.: stopwords: Vietnamese. GitHub (2017)
[3] Trong Hop Do, Slide môn Big Data - Spark for streaming data (2021)
[4] Khanh Q. Tran, An T. Nguyen, Phu Gia Hoang, Canh Duc Luu, Trong-Hop Do and Kiet Van Nguyen, Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data (2022)