# Phát hiện bình luận xúc phạm ngôn ngữ tiếng Anh và tiếng Việt trên dữ liệu trực tuyến từ Twitter

Đặng Văn Nhân[18521172], Võ Hồng Thiên[18521432], Nguyễn Thế Thành[18521411], and Phạm Vũ Kiều Tiên[1851490]

University of Information Technology
Vietnam National University, Ho Chi Minh City, Vietnam

**Tóm tắt nội dung** Twitter được coi là một mạng xã hội nổi tiếng thu hút các tổ chức, cá nhân bày tỏ quan điểm của mình về các vấn đề chính trị, kinh tế và xã hội. Song, hiện tượng công kích người dùng bằng những lời bình luận mang tính công kích, xúc phạm là một trong những vấn đề nan giải về văn hóa ứng xử trên mạng xã hội trong những năm gần đây. Những bình luận như vậy rất quan trọng để những người quản trị mạng xã hội có thể ẩn hoặc xóa nó đi, nhằm tránh gây ra các hậu quả tiêu cực cho người dùng bị ảnh hưởng. Trong bài báo cáo này, chúng tôi đề xuất một mô hình real-time tự động phân loại các bình luận của người dùng trên Twitter sau mỗi Tweet được đăng tải. Mô hình được xây dựng bằng cách áp dụng các phương pháp máy học Logistic Regression, Decision Tree, Naive Bayes trên hai bộ dữ liệu tiếng Anh và tiếng Việt. Hiệu suất của các mô hình phân loại được đo bằng F1 score. Kết quả cho thấy rằng mô hình Naive Bayes của chúng tôi đạt F1 score cao nhất ở cả hai bộ dữ liệu được đề cập.

**Keywords:** Hate Speech Detection · Streaming Data · Spark Machine Learning. · Big Data

## 1 Giới thiệu

### 1.1 Thực trạng

Internet đã định hình cách chúng ta nhận thức về thế giới. Truyền thông xã hội cũng là một phần của Internet, được biểu hiện dưới nhiều dạng, nền tảng: game trực tuyến, ứng dụng hẹn hò, trang báo điện tử, mạng xã hội. Các mạng xã hội khác nhau thường nhắm đến các đối tượng, mục đích khác nhau: chia sẻ ý kiến (Twitter, Facebook), chia sẻ hình ảnh, video (Instagram, Youtube), công việc (LinkedIn). Mặc dù vậy, tất cả các mạng xã hội đều có một điểm chung đó là kết nối mọi người lại với nhau. Do đó, sự phổ biến của mạng xã hội đã vượt ngoài sức tưởng tượng của chúng ta và đạt được hơn 3 tỉ lượt hoạt động trên từng tháng trong năm 2021.

Chính vì thế, mạng xã hội là một nguồn dữ liệu cực lớn cho các nhà khoa học dữ liệu có thể phân tích, nghiên cứu. Ví dụ điển hình là Twitter, một mạng xã

hội với nguồn dữ liệu khổng lồ và là một trong những nguồn dữ liệu quan trọng nhất với các nhà nghiên cứu. Twitter được biết đến như là một real-time blog, nơi mà cộng đồng đăng tải các tin tức trước cả các trang báo điện tử, ngoài ra còn có lượng bài đăng, bình luận cá nhân cực lớn. Các bài đăng được giới hạn 280 từ và đạt trung bình 500 triệu bài đăng trong các sự kiện lớn. Tuy nhiên, trong những năm gần đây, số lượng người dùng sử dụng các ngôn từ xúc phạm và độc hại ngày một gia tăng, dẫn đến hậu quả là những ảnh hưởng xấu đến người dùng khác. Điều này tạo ra một thách thức lớn cho các tổ chức chính phủ và phi chính phủ trong việc quản lý, xử lý các tương tác độc hại trên nền tảng này.

### 1.2 Giới thiệu bài toán

Thực trạng trên có ở khắp các mạng xã hội, đặc biệt là Twitter. Hiện nay Twitter là nơi kết nối giao tiếp giữa mọi người trên khắp thế giới trong đó có Việt Nam chúng ta. Các bài đăng, tin tức trên mạng xã hội bao gồm tất cả các thể loại, vấn đề trong đời sống và đem lại một lượng tương tác khổng lồ. Việc thực hiện tương tác trên Twitter rất dễ dàng và tiện lợi mà không qua các bước kiểm duyệt nào. Vì Twitter không thực hiện kiểm duyệt nên thực trạng ấy được thể hiện một cách rõ ràng nhất. Vì vậy, ở trong bài báo cáo này, chúng tôi muốn giới thiệu một hệ thống máy học real-time để nhận diện các bình luận xúc phạm trên Twitter với ngôn ngữ tiếng Anh và tiếng Việt (hai ngôn ngữ phù hợp cho tác vụ này). Hệ thống này sẽ thu thập các bình luận trên Twitter theo thời gian thực và sau đó phân loại chúng có phải là bình luận xúc phạm hay không. Qua đó có thể giúp các tổ chức tiến hành phát hiện và xử lý các cá nhân đăng tải bình luận này hoặc đơn giản hơn là xóa các bình luận xấu đi.

## 2 Bộ dữ liệu

### 2.1 Train test data

Chúng tôi sử dụng bộ dữ liệu UIT-ViHSD để huấn luyện và thử nghiệm mô hình Tiếng Việt và Twitter-HSD để huấn luyện mô hình Tiếng Anh.

Bộ dữ liệu UIT-ViHSD bao gồm hơn 30,000 bình luận Tiếng Việt trên mạng xã hội, mỗi bình luận thuộc 1 trong 3 nhãn: CLEAN, OFFENSIVE hoặc HATE.

Bộ dữ liệu Twitter-HSD bao gồm 24,784 bình luận Tiếng Anh được thu thập từ mạng xã hội Twitter, mỗi bình luận thuộc 1 trong 3 nhãn: CLEAN, OFFENSIVE hoặc HATE.

**Bảng 1.** Số lượng các nhãn trong mỗi bộ dữ liệu.

| | Bình thường | Phản cảm | Ghét |
|---|---|---|---|
| UIT-ViHSD | 27624 | 2262 | 3514 |
| Twitter-HSD | 4181 | 19190 | 1430 |

**Dữ liệu trực tuyến:** Để áp dụng vào thực tế, chúng tôi dùng Spark Streaming để lấy bình luận trực tuyến từ Twitter cho việc phát hiện nội dung mang tính xúc phạm.

# 3 Phương pháp

## 3.1 Tiền xử lý và trích xuất feature

Trước khi tiến hành huấn luyện mô hình, cần phải xử lý văn bản. Với mỗi bình luận chúng tôi làm sạch bằng cách loại bỏ các stop words, ký tự đặc biệt, thẻ html và link.

Sau đó tiến hành lọc bộ dữ liệu bằng cách loại bỏ các bình luận rỗng.

Dữ liệu sau khi được xử lý sẽ được whitespace tokenize.

Tiếp đến, chúng tôi sử dụng TF-IDF (Term Frequency – Inverse Document Frequency)[1] để tạo bộ từ điển. Ta có thể tính giá trị **TF-IDF = TF x IDF**

Với:

$$tf(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

- tf(t,d): tần suất xuất hiện của từ t trong văn bản d.
- f(t,d): Số lần xuất hiện của từ t trong văn bản d.
- $\max\{\mathrm{f}(w,d) : w \in d\}$: Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

- idf(t, D): giá trị idf của từ t trong tập văn bản.
- |D|: Tổng số văn bản trong tập D.
- $|\{d \in D : t \in d\}|$ : thể hiện số văn bản trong tập D có chứa từ t.

## 3.2 Mô hình huấn luyện

### 3.2.1 Logistic Regression [2]

Logistic Regression là một mô hình máy học hồi quy nhưng lại được áp dụng chủ yếu vào các bài toán phân loại. Các lý do chúng tôi chọn mô hình này là:

- Logistic Regression cũng được sử dụng phổ biến trong bài toán Hate Speech Detection
- Các bình luận đều độc lập. Vậy nên việc áp dụng Logistic Regression sẽ đem lại kết quả tốt.
- Loss function của Logistic Regression là một hàm xác suất giúp dễ dàng phân loại dữ liệu.
- Kiến trúc thuật toán không quá phức tạp, việc huấn luyện mô hình không tốn nhiều thời gian.

**3.2.2 Multinomial Naive Bayes** [3]
Bộ phân loại Naive Bayes (NBC) được lựa chọn sử dụng vì:

- Bộ phân loại Naive Bayes (NBC) thường được sử dụng trong các bài toán phân loại văn bản.
- NBC có thời gian huấn luyện và kiểm tra rất nhanh. Điều này có được là do giả sử về tính độc lập giữa các thành phần.
- NBC có thể hoạt động với các vector đặc trưng ở dạng rời rạc (sử dụng multinomial hoặc Bernoulli). Sự độc lập giữa các đặc trưng khiến NBC có khả năng này.
- Làm mềm Laplace được sử dụng để tránh trường hợp một từ trong tập kiểm tra chưa xuất hiện trong tập huấn luyện.

**3.2.3 Decision Tree** Decision Tree được chúng tôi lựa chọn vì việc chuẩn bị dữ liệu cho Decision Tree rất cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi chuẩn hóa dữ liệu, cần tạo các biến phụ (dummy variable) và loại bỏ các giá trị rỗng. Ngoài ra, cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng máy tính cá nhân để phân tích các lượng dữ liệu lớn trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.

### 3.3 Spark Streaming

Spark Streaming là một extension của Spark API cho phép mở rộng, thông lượng cao, xử lý dữ liệu trực tuyến. Dữ liệu có thể được nạp vào từ nhiều nguồn như Kafka, Kinesis, hoặc TCP sockets và có thể được xử lý bằng nhiều thuật toán phức tạp.

Chúng tôi đã sử dụng socket làm cổng trao đổi dữ liệu trong dự án này. Ngay sau khi nhận bình luận từ Twitter, chúng tôi sẽ thực hiện làm sạch và dự đoán bình luận đã nhận có tính công kích, xúc phạm hay không.

## 4 Kết quả thực nghiệm

**Bảng 2.** So sánh kết quả giữa ba mô hình máy học.

| | UIT-ViHSD(F1) | Tweet-HSD (F1) |
|---|---|---|
| Naive Bayes | 82.99% | 84.15% |
| Logistic Regression | 81.11% | 79.9% |
| Decision Tree | 75.35% | 68.1% |

Kết quả chạy trên bộ dữ liệu test đem lại kết quả khá tốt. Mô hình đã dự đoán đúng phần lớn về các bình luân có xúc phạm hay là không.

**Bảng 3.** So sánh kết quả dự đoán từ mô hình và nhãn thực của bộ dữ liệu tiếng Anh.

| Comment | Prediction | Truth Label |
|---|---|---|
| Got a new desk for my tablet. Its to big for my main desk | CLEAN | CLEAN |
| I want same news about you | CLEAN | CLEAN |
| I'm loving you from a distance but the road is getting longer | OFFENSIVE | OFFENSIVE |
| Your honor I was just being a silly little goose | HATE | OFFENSIVE |
| What a disgraceful article Polly Hudson | HATE | HATE |

**Bảng 4.** So sánh kết quả dự đoán từ mô hình và nhãn thực của bộ dữ liệu tiếng Việt.

| Comment | Prediction | Truth Label |
|---|---|---|
| Đẹp quá đi à | CLEAN | CLEAN |
| Cảm thấy may mắn vì biết đến chị mà giỏi tiếng anh hẳn lên | CLEAN | CLEAN |
| Tôi cảm thấy rất thất vọng về bản thân mình | OFFENSIVE | OFFENSIVE |
| Tôi không thể tin là có người làm một điều ngu ngốc đến như vậy | HATE | OFFENSIVE |
| Nếu tôi nói anh ngu như chó thì tôi đang xúc phạm loài chó đó | HATE | HATE |

# 5 Kết luận

## 5.1 Kết quả đạt được

- Chọn được 2 bộ dữ liệu phù hợp cho bài toán Hate Speech Detection cho tiếng Anh và tiếng Việt.
- Xin được API từ Twitter.
- Hiểu và áp dụng được streaming dữ liệu theo thời gian thực từ mạng xã hội Twitter.
- Huấn luyện được đa dạng các mô hình máy học cho bài toán.
- Độ chính xác khả quan (trên 82% ở cả 2 bộ dữ liệu).

## 5.2 Hạn chế

- Chỉ mới áp dụng trên Twitter, bình luận tiếng Việt trên nền tảng này chưa phổ biến như Facebook hay Youtube.
- Chưa áp dụng được các mô hình học sâu với Spark.

## 5.3 Hướng phát triển

- Mở rộng mô hình áp dụng cho đa nền tảng, mạng xã hội.
- Tìm hiểu sử dụng BigDL cho các mô hình học sâu.

# Tài liệu

1. Sundaram, Varun Ahmed, Saad Muqtadeer, Shaik Reddy, R., `https://www.researchgate.net/publication/350080855_Emotion_Analysis_in_Text_using_TF-IDF`. Truy cập lần cuối 25/01/2022.

2. Backhaus, Klaus Erichson, Bernd Gensler, Sonja Weiber, Rolf Weiber, Thomas, https://www.researchgate.net/publication/355261513_Logistic_Regression. Truy cập lần cuối 25/01/2022.
3. Kibriya, Ashraf Frank, E. Pfahringer, Bernhard Holmes, Geoffrey, https://www.researchgate.net/publication/287228965_Multinomial_naive_Bayes_for_text_categorization_revisited. Truy cập lần cuối 26/01/2022.