## 0. Tóm tắt nội dung Mạng xã hội không còn là thứ quá xa lạ đối với mỗi người. Tuy nhiên vì quá lạm dụng mạng xã hội dẫn đến ngày càng có nhiều bình luận độc hại gây ảnh hưởng đến tâm lý người dùng. Vì vậy, viêc quan tâm đến tích độc hại của bình luận là vô cùng cần thiết. Do đó chúng em xây dựng bài toán Nhận diện tính độc hại của bình luận tiếng việt trên mạng xã hội trên bộ dữ liệu UIT-ViCTSD của Nguyễn Thành Luân và cộng sự với kích thước 10.000 bình luận. Sau đó, chúng em tiến hành tiền xử lý dữ liệu và thử nghiệm, đánh giá bộ dữ liệu trên các phương pháp máy học truyền thống. Kết quả thu được thể hiện tốt nhất cho bài toán với F1-macro là 68.94%. Tuy nhiên, kết quả hiện tại chưa phải là tốt nhất, do có sự chênh lệch rất lớn giữa nhãn độc hại và không độc hại trong bộ dữ liệu. Trong tương lai chúng em sẽ hướng tới việc cân bằng lại dữ liệu cũng như thực nghiệm bằng các mô hình học sâu để tìm ra một phương pháp có thể đưa ra các kết quả tốt nhất và từ đó hướng tới việc xây dựng một hệ thống lọc bình luận có tính độc hại để xây dựng một không gian mạng tích cực hơn.

## 

## 1. Giới thiệu

## Hiện nay khi internet đã được phủ sóng rộng rãi tới khắp mọi miền đất nước thì mạng xã hội không còn là thứ quá xa lạ đối với mỗi người. Đây là nơi kết nối con người khắp nơi trên thế giới, xóa nhòa khoảng cách về không gian nhờ tốc độ nhanh chóng, tiện lời. Nhưng chính vì quá lạm dụng mạng xã hội dẫn đến việc ngày càng có nhiều bình luận độc hại gây ảnh hưởng không ít đến tâm lý người dùng. Vì vậy, viêc quan tâm đến tích độc hại của bình luận cũng góp phần xây dựng, phát triển các cuộc thảo luận trên không gian mạng trở nên tích cực hơn. Chính vì thế, chúng em kế thừa bộ dữ liệu tiếng việt UIT-ViCTSD [1] (**V**ietnamese **C**onstructive and **T**oxic **S**peech **D**etection dataset) với kích thước 10,000 bình luận. Sau đó, chúng em tiến hành thử nghiệm và đánh giá bộ dữ liệu trên các phương pháp máy học truyền thống. Từ bộ dữ liệu, chúng em phát triển bài toán Nhận diện tính độc hại của bình luận tiếng việt trên mạng xã hội.

## 2. Giới thiệu bộ dữ liệu

## Phần lớn các bình luận tiếng việt trên mạng xã hội thường theo xu hướng teencode, không theo cấu trúc ngữ pháp tiếng việt, có nhiều kí tự đặc biêt, sai chính tả, trong một câu có thể đan xen nhiều loại ngôn ngữ như: tiếng Việt, tiếng Anh, tiếng Hàn,… Với sự đa dạng về mặt ngữ nghĩa của tiếng Việt thì tùy vào từng ngữ cảnh khác nhau mà ý nghĩa khác nhau. Chính vì thế để xây dựng một bộ dữ liệu tốt sẽ gặp rất nhiều khó khăn. Vì vậy chúng em kế thừa bộ dữ liệu UIT – ViCTSD của Nguyễn Thành Luân và cộng sự với kích thước 10.000 bình luận và được chia thành 10 lĩnh vực, bao gồm entertainment, education, science, business, cars, law, health, world, sports, và news.

## Bộ dữ liệu có các thuộc tính như id, comment, constructiveness, toxicity, title, topic. Nhưng trong đồ án của chúng em chỉ cần quan tâm đến thuộc tính comment và toxicity, nên đã loại bỏ những thuộc tính không cần thiết ra khỏi bộ dữ liệu. Sau đó thay đổi nhãn của thuộc tính toxicity (non\_toxic, toxic) thành (0,1), để tiện cho việc chạy mô hình.

## 

## a. Định nghĩa bài toán

## - Đầu vào: Một bình luận tiếng việt của người dùng.

## - Đầu ra: Một trong 2 nhãn độc hại (Toxic) và Không độc hại (Non - Toxic).

## b. Định nghĩa nhãn của dữ liệu

## - oxic (Độc hại): Là những bình luận có nội dung châm biếm, chỉ trích; có một sự chế giễu hoặc có thái độ không đồng tình với ý kiến của người khác nhưng thiếu tế nhị, thiếu lịch sự.

## - Non - Toxic (Không độc hại): Là những bình luận không có nội dung châm biếm, chỉ trích, chế giễu. Mang cảm xúc trong sáng hoặc không có nhiều ý nghĩa.

## Bảng 1: Ví dụ về những bình mluận độc hại và không độc hại trong bộ dữ liệu. 1= Độc hại và 0 = Không độc hại

| STT | Bình Luận | Nhãn |
| --- | --- | --- |
| 1 | Thương các bé quá! Các con còn quá nhỏ mà đã phải sớm rời xa cha mẹ! Xin chia buồn cùng các gia đình các con, cầu mong các con thật vui ở thế giới mới! | 0 |
| 2 | Học đại học còn có mấy bạn vô ý thức, quan trọng con người thôi | 1 |
| 3 | Thắng là vui rồi! Đội bóng lớn phải là đội không những phải biết vùi dập đối thủ trong một ngày đẹp trời mà còn phải biết kiếm được 3 điểm trong một ngày mình chơi không tốt! Chúc mừng gooner, tôi tin là năm nay với những con người này, Arsenal sẽ tạo ra một điều đặc biệt! | 0 |
| 4 | Nghe các bạn online phán mất mệt..hông ai sát đánh bắt sát sinh lấy gì cho mấy bạn ăn vậy? Chỉ là tai nạn thôi nha mấy thánh online | 1 |

## c. Khảo sát tập dữ liệu

## Để có cái nhìn tổng quan về bộ dữ liệu, chúng em tiến hành phân tích số lượng nhãn. Hình 1 thể hiện biểu đồ phân bố nhãn trong bộ dữ liệu. Chúng em nhận thấy rằng có sự mất cân đối trong bộ dữ liệu. Những bình luận có tính độc hại ít hơn bình luận không có tính độc hại. Do đó, chúng em sẽ tiến hành cân bằng bộ dữ liệu trong tương lai.

Hình 1: Biểu đồ phân bố nhãn trong Hình 2: Biểu đồ phân bố số lượng từ trong

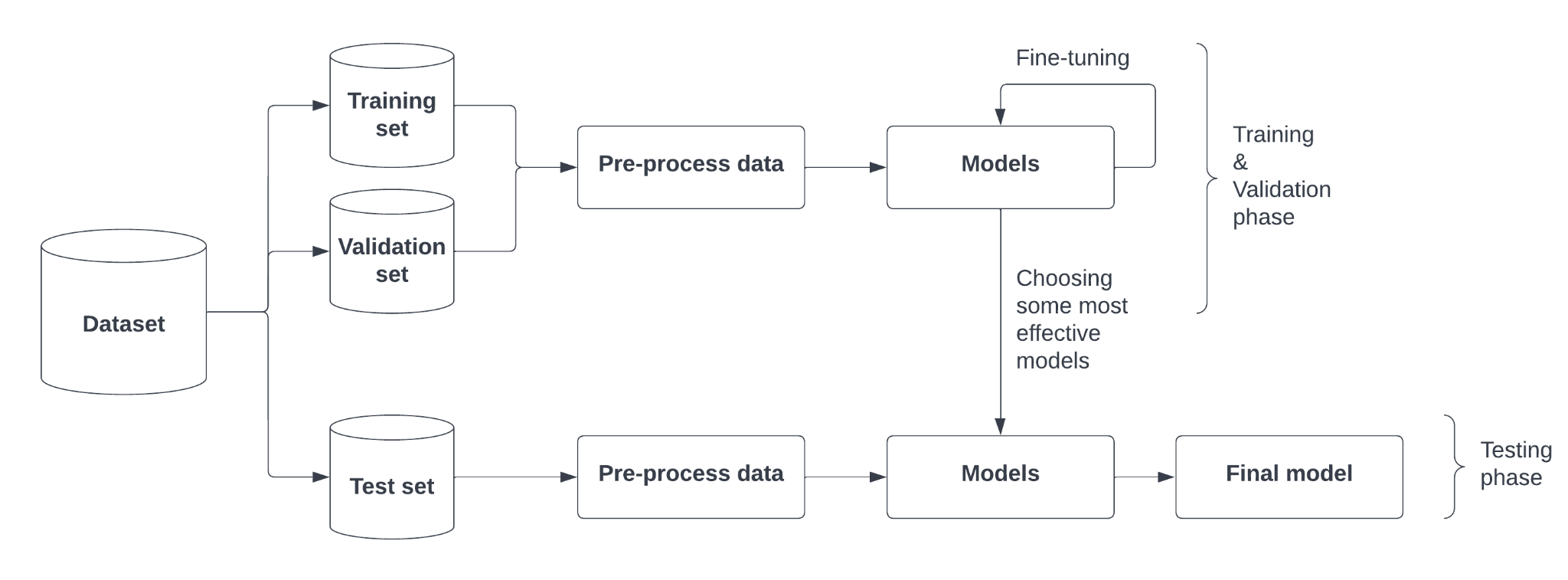
bộ dữ liệu. câu của mỗi nhãn.

Ngoài ra, chúng em còn phân tích thêm số lượng từ trong câu của mỗi nhãn. Kết quả được thể hiện trong hình 2, những bình luận mang tính không độc hại thường là những bình luận dài và có khác biệt đáng kể so với bình luận độc hại.

## 3.Đề xuất hệ thống

### Tổng quan hệ thống

Trong đồ án này, chúng em đề xuất xây dựng một hệ thống cho việc xác định bình luận tiếng Việt độc hại trên mạng xã hội..



Hình 3: Hệ thống đề xuất cho việc xác định bình luận độc hại trên mạng xã hội.

Với hệ thống này, chúng em sẽ tiền xử lý các set của bộ dữ liệu. Ở pha Huấn luyện và thẩm định (Training & Validation phase), sử dụng Training set để huấn luyện và tối ưu hóa tham số các mô hình học máy với các siêu tham số cụ thể, và được thẩm định bởi Validation set. Sau đó, chúng em sẽ lọc lấy những cặp siêu tham số giúp mô hình đạt được hiệu quả cao nhất.

Và cuối cùng ở pha kiểm định (Testing phase) cho các mô hình được kiểm định lần cuối với Test set, chọn ra 3 bộ siêu tham số đem lại hiệu quả cao nhất với từng mô hình học máy.

### Tiền xử lý dữ liệu

Bộ dữ liệu đã được tiền xử lý từ trước, nhưng chúng em cũng áp dụng các phương pháp tiền xử khác như tách từ (word segmentation), đưa về văn bản viết thường, xóa các ký tự đặc biệt, xóa khoảng trắng dư, xóa stop-word. Trong đó dữ liệu cho bộ stop-word tiếng Việt được lấy từ topic Vietnamese StopWords W2V trên trang web Kaggle, và hàm tách từ được sử dụng từ thư viện underthesea.

Dữ liệu trước và sau tiền xử lý được thể hiện ở bảng {số bảng} sau đây.

Bảng 2: Một số ví dụ trước và sau khi tiền xử lý dữ liệu.

| **Raw** | **Pre-processed** |
| --- | --- |
| Đáp án là không bao giờ có một tam giác như thế nhé! | đáp\_án là không bao\_giờ có một tam\_giác như thế nhé |
| Trong các mỹ nhân trên thì Diểm My trông già dặn hơn nhiều. | trong các mỹ\_nhân trên thì diểm\_my\_trông già\_dặn hơn nhiều |

Sau đó, dữ liệu được mã hóa bằng các encoder như sau:

* CountVectorizer: Được sử dụng để chuyển đổi một bộ sưu tập các tài liệu văn bản thành một vectơ có số lượng thuật ngữ / mã thông báo. Nó cũng cho phép xử lý trước dữ liệu văn bản trước khi tạo biểu diễn vectơ. Chức năng này làm cho nó trở thành một mô-đun biểu diễn tính năng rất linh hoạt cho văn bản.
* TfidfVectorizer: Là một phương thức thống kê được biết đến rộng rãi nhất để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau. Nó thường được sử dụng như một trọng số trong việc khai phá dữ liệu văn bản. TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector (VSM), hoặc thành những vector thưa thớt.

## 4. Thử nghiệm và kết quả

### Các mô hình thử nghiệm

Các mô hình học máy

* Naive Bayes: Là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes, Thuật toán này thuộc nhóm Supervised Learning (Học có giám sát). Chúng em chọn kiểu mô hình Multinomial Naive Bayes, vì mô hình này chủ yếu được sử dụng trong phân loại văn bản. Đặc trưng đầu vào ở đây chính là tần suất xuất hiện của từ trong văn bản đó.
* Logistic Regression: Là một thuật toán học máy cơ bản cho việc phân lớp, đặc biệt là các bài toán phân lớp nhị phân.
* Support Vector Machine (SVM): Đây là một thuật toán khá hiệu quả trong lớp các bài toán phân loại nhị phân và dự báo của học có giám sát. Thuật toán này có ưu điểm là hoạt động tốt đối với những mẫu dữ liệu có kích thước lớn và thường mang lại kết quả vượt trội so với lớp các thuật toán khác trong học có giám sát.

### Cài đặt thử nghiệm

Để tìm được bộ tham số phù hợp nhất với từng mô hình học máy, thì chúng em sử dụng phương pháp GridSearchCV của sklearn cho Logistic Regression và SVM. Ở cả mô hình Logistic Regression và SVM đều có thông số Cross-Validation (CV). Chúng em sẽ thử nghiệm trên thông số CV ∈ [2,5].

Ngoài ra ở phần mã hóa, CountVectorizer và TfidfVectorizer có thông số ngram\_range(lower, upper). Ranh giới dưới (lower) và ranh giới trên (upper) của dải giá trị n cho các từ n-gram hoặc char n-gram khác nhau sẽ được trích xuất. Tất cả các giá trị của n sao cho min\_n <= n <= max\_n sẽ được sử dụng. Chúng em sẽ sử dụng (1,1), (1,2) và (2,2).

Bộ thông số cuối cùng được chúng em trình bày ở bảng phía dưới (Bảng {số bảng}).

Bảng 3: Bộ siêu tham số phù hợp cho mỗi mô hình phân loại bình luận tiếng Việt toxic trên mạng xã hội.

| Mô hình | Tiền xử lý | Bộ mã hóa | GridSearchCV |
| --- | --- | --- | --- |
| Naive Bayes | Không | CountVectorizer: ngram\_range(1,1) | Không |
| Logistic Regression | Có | TfidfVectorizer: ngram\_range(1,1) | CV: 3  C: 1000.0  multi\_class: multinomial  penalty: l2 |
| SVM | Không | CountVectorizer: ngram\_range(1,1) | CV: 3  C: 10  gamma: 0.01  kernel: rbf |

### Kết quả thử nghiệm

Chúng em chạy thử tất cả các trường hợp, kiểm thử trên dev set, sau đó thu thập kết quả dựa trên các thông số F1-macro, accuracy, precision, recall (%). Tuy nhiên, chúng em chú trọng thông số F1-macro bởi vì nó là điều hòa của precision và recall, có thể thể hiện được mức độ hiệu quả của mô hình.

Sau khi chạy thử, chúng em chọn ra các thông số mà mô hình đem lại điểm F1-macro tốt nhất, cụ thể là trên 63%. Tiếp theo, chúng em chạy thử mô hình với những thông số đó trên test set. Và chọn ra bộ thông số đem lại hiệu quả nhất cho từng mô hình học máy.

.

Bảng 4: Kết quả thử nghiệm các mô hình

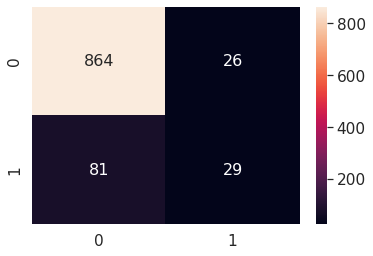
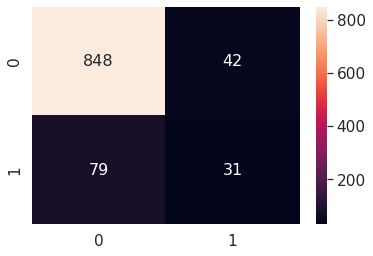
| Model | F1-macro | accuracy | precision | recall |
| --- | --- | --- | --- | --- |
| Logistic Regression | **68.9410916** | 90.7 | 66.03773585 | 31.81818182 |
| Naive Bayes | 63.61022643 | 87.9 | 42.46575342 | 28.18181818 |
| SVM | 64.66022624 | 89.3 | 52.72727273 | 26.36363636 |

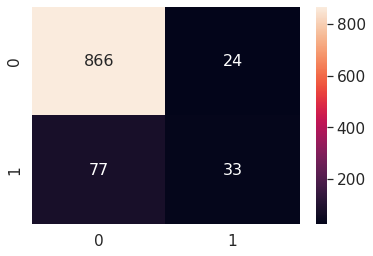
Với kết quả thử nghiệm như trên, mô hình của chúng em thu nhận được kết quả thể hiện tốt nhất cho bài toán phân loại tính độc hại của bình luận tiếng Việt trên mạng xã hội với F1-macro là 68.94%. Tuy nhiên, kết quả hiện tại có lẽ chưa phải là tốt nhất, do có sự chênh lệch rất lớn giữa nhãn toxic và non\_toxic trong bộ dữ liệu, dựa vào bảng {số bảng phân tích số lượng nhãn của dữ liệu}. Và điều đó thể hiện rằng trong bộ dữ liệu vẫn tồn tại những dữ liệu nhiễu khiến cho mô hình phân lớp bị ảnh hưởng trực tiếp.

### d.Phân tích lỗi

Để có thể dễ dàng phân tích lỗi, chúng em tính toán ma trận nhầm lẫn. Bảng {số bảng} thể hiện ma trận nhầm lẫn cho các mô hình học máy bài toán phân loại bình luận tiếng Việt toxic trên mạng xã hội.

Chúng em thấy được rằng ở cả 3 mô hình, số lượng dự đoán đúng ở nhãn 0 (non\_toxic) trên toàn bộ nhãn 0 thực tế cực kì cao (866 đúng, 24 sai), tương đương tỉ lệ 97% cho mô hình Logistic Regression (có F1-macro cao nhất). Và tỉ lệ dự đoán đúng của nhãn 1 là 30% cho mô hình Logistic Regression. Thể hiện rằng mô hình dự đoán nhãn toxic tương đối tệ. Các tỉ lệ cũng xấp xĩ như vậy cho 2 mô hình còn lại. Ma trận nhầm lẫn cho thấy được hiệu quả của mô hình dự đoán bình luận tiếng Việt toxic chưa được quá cao. Và lý do đó phần lớn được thể hiện bởi sự chênh lệch quá lớn giữa nhãn toxic và non\_toxic. Nguồn gốc của bộ dữ liệu UIT-ViCTSD được lấy từ trang mạng xã hội báo chí VnExpress, là một nguồn chính thống, nơi mà đã được kiểm duyệt trước khi được đăng tải lên phần bình luận của forum VnExpress. Vì vậy mà số lượng bình luận toxic chắc chắn sẽ ít hơn rất nhiều so với lượng bình luận non\_toxic. Mô hình học máy từ đó mà cũng cũng sẽ không được huấn luyện tốt cho việc phân loại bình luận toxic.





Hình 4: Ma trận nhầm lẫn Hình 5: Ma trận nhầm lẫn của Hình 6: Ma trận nhầm lẫn của mô hình Naive Bayes mô hình Logistic Regression của mô hình SVM

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Từ các thí nghiệm ở chương 4 ta có thể thấy mô hình Logistic Regression là tối ưu nhất đối với bài toán Nhận diện tính độc hại của bình luận tiếng Việt trên mạng xã hội. Trong tương lai gần chúng em sẽ hướng tới việc cân bằng lại dữ liệu cũng như thực nghiệm bằng các mô hình học sâu để tìm ra một phương pháp có thể đưa ra các kết quả tốt nhất và từ đó hướng tới việc xây dựng một hệ thống lọc bình luận có tính độc hại để xây dựng một không gian mạng tích cực hơn.

Bảng phân công công việc:

-Trần Hoàng Anh: Xây dựng source code, demo. Thống kê và trực quan hóa kết quả mô hình.

-Phạm Tiến Dương: Thống kê và phân tích bộ dữ liệu. Xây dựng slide thuyết trình.

-Trương Phước Bảo Khanh: Viết báo cáo đồ án, phân tích bộ dữ liệu.

-Nguyễn Nhật Hoàng: Chuẩn bị nội dung cho thuyết trình báo cáo đồ án môn học, viết báo cáo đồ án.

-Cả nhóm: Chạy thử và thu nhận kết quả huấn luyện mô hình. Họp bàn đưa ra quy trình làm việc cũng như rút ra các kết luận và phân tích lỗi.