NHẬN DIỆN VÀ PHÂN LOẠI GIẬT GÂN TRÊN TIÊU ĐỀ CỦA BÀI BÁO ĐIỆN TỬ DỰA TRÊN CÁC ĐẶC TRƯNG VỀ NGỮ NGHĨA

Phạm Bá Thuận
Khoa Khoa học và Kĩ thuật thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
22521447@gm.uit.edu.vn

Thái Bình Dương
Khoa Hệ thống thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
23520356@gm.uit.edu.vn

Trương Tất Quang Vinh
Khoa Khoa học và Kĩ thuật thông tin
Trường Đại học Công nghệ thông tin
Thành phố Hồ Chí Minh, Việt Nam
22521683@gm.uit.edu.vn

Tóm tắt*—Clickbait (hay giật gân) là hình thức lôi kéo người đọc hoặc người xem bằng các thủ pháp từ ngữ, nghệ thuật, nội dung,... Trong đề tài này, chúng em sẽ thực hiện phân loại các mức độ giật gân trên tiêu đề của các bài báo điện tử, với dữ liệu là 3908 bài báo tiếng Việt được thu thập thủ công từ nhiều nguồn khác nhau, trong đó 2500 dòng dữ liệu được sử dụng để huấn luyện, số còn lại sẽ được chia thành tập validation và tập test. Sau khi thực hiện gán nhãn dựa trên bốn yếu tố giật gân chủ yếu và tiến hành gộp các yếu tố lại để tính thành nhãn chung, nhóm thực hiện huấn luyện trên mô hình hồi quy Logistic và mô hình SVM, bao gồm các trường hợp sử dụng hàm mất mát SGD, với ba phương pháp: dùng mô hình ngôn ngữ để xử lý (BERT), dùng kỹ thuật TF-IDF và dùng chỉ số LIX, RIX. Kết quả cho thấy phương pháp TF-IDF kết hợp với mô hình SVM mặc dù có điểm Accuracy cao hơn các phương pháp còn lại, nhưng lại tiềm ẩn rủi ro bị quá khớp dữ liệu do chênh lệch điểm giữa tập huấn luyện và tập kiểm thử là khá lớn. Trong khi đó, nếu sử dụng mô hình BERT để xử lý văn bản thì độ chính xác trên tập kiểm thử khi dùng trên nhiều phương pháp máy học là đồng nhất.

Từ khoá—Machine Learning, Support Vector Machine, hồi quy Logistic, Phân loại, Báo điện tử, Đặc trưng ngữ nghĩa

I. GIỚI THIỆU BÀI TOÁN

Với sự phát triển của công nghệ, các bài báo in ấn đang dần được thay thế bằng các trang báo điện tử, như là một phần tất yếu. So với báo giấy truyền thống, báo điện tử có ưu điểm là người dùng dễ dàng tiếp cận thông tin hơn, chỉ bằng một cú nhấn từ các thiết bị di động (điện thoại, tablet, máy tính,…). Ngoài ra, khả năng cập nhật tin tức liên tục theo từng giây phút, và đặc biệt là khả năng truy cập không mất phí cũng là một trong những điều thú vị.

Tại Việt Nam, tính theo số liệu của năm 2016, có gần 1800 trang tin điện tử khác nhau, kể cả nguồn chính thống và không chính thống, chưa kể các trang thông tin trên các nền tảng mạng xã hội. Chính vì số lượng nguồn thông tin lớn nên điều này đã tạo nên sự cạnh tranh giữa các trang thông tin trong việc lôi kéo người đọc [1]. Một trong những phương pháp phổ biến để lôi kéo người đọc trong lĩnh vực báo chí là tạo nên các tiêu đề giật gân ("tít gợi", hay "clickbait" theo thuật ngữ tiếng Anh), sử dụng các từ ngữ hoặc đánh mạnh vào tâm lý người đọc. Tuy nhiên, việc lạm dụng phương pháp này đã tạo nên nhiều bất cập, trong đó không thể không kể đến việc tạo ra các tiêu đề mang tính sai lệch ("tít láo"), gây nhầm lẫn hoặc hoang mang cho người đọc. Chưa kể, một số trang báo không chính thống có thể sẽ kèm các liên kết hoặc nội dung độc hại, gây ra các vấn đề bảo mật và an ninh mạng, và có thể ảnh hưởng đến các vấn đề trật tự xã hội, kích động bạo lực.

Do đó, nhóm chúng em quyết định lựa chọn đề tài "Nhận diện và phân loại giật gân trên tiêu đề của bài báo điện tử dựa trên các đặc trưng ngữ nghĩa". Nhóm sẽ thực hiện xây dựng dữ liệu gồm các tiêu đề, chủ đề, nguồn và tóm tắt của các bài báo điện tử. Sau đó, thực hiện gán nhãn dữ liệu dựa trên một thang quy ước gồm bốn yếu tố: nội dung, từ ngữ, dấu câu, số từ; với giá trị gán nhãn là 0 và 1. Cuối cùng, thực hiện huấn luyện dữ liệu đã gán nhãn với hai phương pháp học máy là hồi quy Logistic và Support Vector Machine (SVM). Nghiên cứu này có ý nghĩa vô cùng lớn trong việc phân loại và nâng cao độ tin cậy của các tin tức trên mạng, giúp người đọc phòng tránh những sự việc tiêu cực hoặc bị "dắt mũi" bởi những thông tin sai lệch.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong lĩnh vực Khoa học máy tính, đã có khá nhiều nghiên cứu về vấn đề tiêu đề giật gân. Một số nghiên cứu có thể nhắc đến dưới đây:

Prakhar Biyani và các cộng sự [2] đã nghiên cứu các dạng clickbait trong các tiêu đề trang web tin tức dựa trên nội dung, độ tương đồng của tiêu đề so với nội dung chi tiết, đồng thời tính trang trọng và khả năng vào vấn đề trực tiếp (Informality and Forward Reference) bằng số điểm Coleman-Liau (CLScore), chỉ số RIX, LIX và thước đo phổ thông (fmeasure), và đưa ra 8 dạng giật gân khác nhau. Cuối cùng, nhóm nghiên cứu này đã thực hiện xây dựng, gán nhãn dữ liệu và huấn luyện trên mô hình Gradient Boosted Decision Trees (GBDT).

Một nghiên cứu khác của Shiva Ram Dam cùng các cộng sự [3] sử dụng phương pháp Support Vector Machine và Random Forest. Cả hai mô hình này sử dụng thang đo Cosine Similarity và chỉ số TF-IDF để so sánh mức độ tương đồng giữa tiêu đề và nội dung tin tức.

Ngoài ra, xét theo hướng nghiên cứu đặc trưng ngữ nghĩa chúng ta có nhóm của Bronakowski [4]. Trong nghiên cứu này, nhóm tác giả đã chia ra đến 30 đặc trưng cốt lõi theo 6 nhóm: số lượng từ ngữ, tính chất vào vấn đề trực tiếp, khả năng đọc hiểu (mức độ trang trọng), cùng đặc trưng từ khoá, dấu câu, chữ số,… và thực hiện phân loại bằng 6 mô hình khác nhau, với độ chính xác dao động trong khoảng 0,96 đến 0,976. Nghiên cứu của Ahmadi và các cộng sự cũng có cách thực hiện tương tự [5], tuy nhiên nhóm có sử dụng thêm tóm tắt văn bản bằng mô hình Text-to-text Transfer Transformer.

Mặc dù vậy, chưa có một nghiên cứu cụ thể về giật gân trong tiêu đề các bài báo tiếng Việt. Vì thế, trong đồ án này, nhóm sẽ sử dụng các đặc trưng trong nghiên cứu về tiêu đề bài báo của các ngôn ngữ khác (ví dụ như tiếng Anh).

III. DỮ LIỆU

A. Tổng quan

Trong mục này, chúng em sẽ đề cập về vấn đề thu thập, gắn nhãn và xây dựng một bộ dữ liệu để thực hiện bài toán, cũng như trình bày khó khăn trong việc gắn nhãn dữ liệu và đánh giá độ đồng thuận.

Hình 1: Ví dụ về các yếu tố có trong bài báo, gồm tiêu đề, chủ đề và tóm tắt

B. Thu thập và xây dựng dữ liệu

Thông thường, một bài báo điện tử sẽ gồm có các yếu tố như sau:

- Tiêu đề bài báo: thường có cỡ chữ to, đậm, dễ chú ý.
- Chủ đề bài báo đang hướng đến (Topic): thường đặt theo dạng mục, đặt ngay trên tiêu đề.
- Tóm tắt hoặc tổng quan bài viết (Summary): thường sẽ in đậm và đặt ngay đầu bài báo, dưới phần tiêu đề.

Ví dụ minh hoạ về các yếu tố đề cập ở trên có thể thấy được ở Hình 1.

Vì thế, nhóm đã thực hiện thu thập các yếu tố trên bằng phương pháp thủ công hoặc crawl bằng phần mềm Octoparse, kết quả ban đầu đã thu được các tiêu đề, chủ đề, tóm tắt của 4200 bài báo với các nguồn đa dạng. Tuy nhiên, chỉ có 3908 dữ liệu được sử dụng, lí do sẽ được nhóm chúng em đề cập ở phần sau.

C. Gán nhãn dữ liệu

Ban đầu, nhóm đã xây dựng gán nhãn thử với 200 dòng dữ liệu dựa trên ba yếu tố chính:

- Nội dung: Mức độ tương đồng về ý nghĩa giữa tiêu đề và tóm tắt.
- Từ ngữ: Sử dụng các từ ngữ giật gân (sốc, ngỡ ngàng, nhất, tin nóng…) hoặc thủ pháp nghệ thuật, ca dao, tục ngữ,…
- Cấu trúc: Tiêu đề có cấu trúc ngữ pháp cơ bản hay không hoặc quá dài hay quá ngắn không,…

Các yếu tố được đánh giá theo thang điểm từ 0 đến 3. Về sau, yếu tố về cấu trúc được loại bỏ, chỉ còn hai yếu tố do cấu trúc không có quá nhiều ý nghĩa về học máy.

Lần gán nhãn đầu tiên được thực hiện trên 8 người với 200 dòng dữ liệu để đo độ đồng thuận. Tuy nhiên, độ đồng thuận giữa những người gán nhãn là không đồng đều khi hệ số Cohen's kappa xét theo từng cặp nằm trong khoảng từ 0,159 đến 0,676, nghĩa là có sự chênh lệch khá lớn. Lý giải cho điều này, nhóm chúng em cho rằng khái niệm "giật gân" với mỗi người sẽ khác nhau, phụ thuộc vào tâm lý của từng người. Việc sử dụng gãn nhãn theo thang mức độ có thể sẽ không phù hợp.

Vì lí do này, nhóm đã sửa đổi lại quy tắc gán nhãn, thay vì gán nhãn theo mức độ thì sẽ chỉ còn là gán nhãn nhị phân với hai giá trị là 0 (không có yếu tố đó) hoặc 1 (có yếu tố đó). Ngoài nội dung và từ ngữ, nhóm cũng đã thêm hai đặc trưng khác để gán nhãn là: số từ, và dấu câu. Số lượng thành viên tham gia gán nhãn giảm còn 3 người, và sau bốn lần gán nhãn thử với 250 dòng dữ liệu, nhóm đã thực hiện các sửa đổi và cuối cùng đã rút ra được chi tiết về quy tắc gán nhãn (được trình bày trong Bảng *I*).

Tuy nhiên, một vấn đề phát sinh khác đó chính là tiêu đề giật gân xét theo từng yếu tố đặc trưng là quá ít. Do đó, chúng em đã tính đến phương án là kết quả sẽ được tổng hợp lại thành một cột giá trị là tổng (Total). Nếu tiêu đề có chứa bất kì một đặc trưng nào, giá trị Total sẽ bằng 1, ngược lại là 0.

Sau khi đã thực hiện gán nhãn, có 1954 trên tổng số 4200 dòng dữ liệu thu thập được có giá trị Total bằng 1. Để đảm bảo ý nghĩa về học máy, nhóm đã thực hiện loại bỏ ngẫu nhiên một số tiêu đề có giá trị Total bằng 0, sao cho tổng số giá trị giữa 0 và 1 ở cột Total là bằng nhau. Như vậy, có tổng 3908 bài báo được sử dụng cho bài toán này.

D. Xử lý dữ liệu văn bản

Để xử lý văn bản có trong tiêu đề và tóm tắt, nhóm đã thực hiện thí nghiệm với ba phương pháp phân tích và xử lý ngôn ngữ tự nhiên gồm: Vectơ hoá bằng kỹ thuật TF-IDF [6], tính toán chỉ số LIX và RIX [7], và chuyển đổi văn bản thành dạng token bằng mô hình BERT [8].

Bảng I: Quy ước gán nhãn bộ dữ liệu theo đặc trưng

Đặc trưng / Giá trị	Nội dung	Từ ngữ	Dấu câu	Số từ
0	Tiêu đề có nội dung giống hoặc có liên kết với phần tóm tắt	Không chứa các từ ngữ giật gân	Tiêu đề không sử dụng các dấu câu đặc biệt để bộc lộ cảm xúc. Tiêu đề sử dụng dấu hai chấm ":" với mục đích trích dẫn hoặc giải thích.	Không sử dụng các con số để lôi kéo người đọc. Chỉ sử dụng con số để biểu thị thứ tự, số năm, đơn vị tiền tệ,…
1	Tiêu đề có nội dung sai lệch hoặc gây ra hiểu lầm so với phần tóm tắt.	Chứa các từ ngữ giật gân: sốc, ngỡ ngàng, đắng lòng,… Sử dụng các từ: ảnh, video, clip,…	Tiêu đề có sử dụng các dấu đặc biệt (?, !), các dạng dấu ngoặc như ngoặc đơn, ngoặc kép, dấu nháy,…	Sử dụng con số để lôi kéo người đọc: *Ví dụ: "3 sai lầm khi ăn chuối mà phần lớn mắc phải".*

TF-IDF (viết tắt của Term Frequency – Inverse Document Frequency) là kỹ thuật thống kê mức độ quan trọng của một từ trong văn bản. Trong đề tài này, TF-IDF dùng để tính toán tần suất xuất hiện các đặc trưng đã đề cập trong phần III.C. Công thức của TF-IDF được thể hiện từ (1) đến (3) như sau:

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{D_j} \tag{1}$$

$$\mathrm{IDF}_i = \lg\left(\frac{D}{N_j}\right) \tag{2}$$

$$\text{TF-IDF}_{ij} = \mathrm{TF}_{ij} * \mathrm{IDF}_i \tag{3}$$

trong đó D_j là văn bản thứ j trong tổng số văn bản D, n_{ij} là số lần xuất hiện của từ thứ i trong văn bản thứ j, và N_i là số lần xuất hiện của từ thứ i trong tất cả các văn bản.

LIX và RIX đều là chỉ số tính toán khả năng đọc hiểu trên một đoạn văn (readability). Chỉ số RIX được tạo nên bởi J.Anderson, dựa trên những nền tảng về công thức LIX của Carl-Hugo Björnsson. Công thức của cả hai chỉ số như sau:

$$\mathrm{LIX} = \frac{W}{S} + \frac{100LW}{W} \tag{4}$$

$$\mathrm{RIX} = \frac{LW}{S} \tag{5}$$

với W là số lượng từ, LW là số lượng cụm từ, và S là số câu. Nói cách khác, RIX là phiên bản giản lược của LIX.

BERT (viết tắt của Bidirectional Encoder Representations from Transformers) là mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer khá phổ biến trong xử lý ngôn ngữ tự nhiên [9]. Mô hình có khả năng chuyển đổi văn bản thành các token là các vectơ phân loại đặc trưng (CLS). Mô hình BERT trong đề tài này sẽ được sử dụng cho biến đổi cả tiêu đề và phần tóm tắt văn bản.

IV. Huấn luyện mô hình

A. Tổng quan

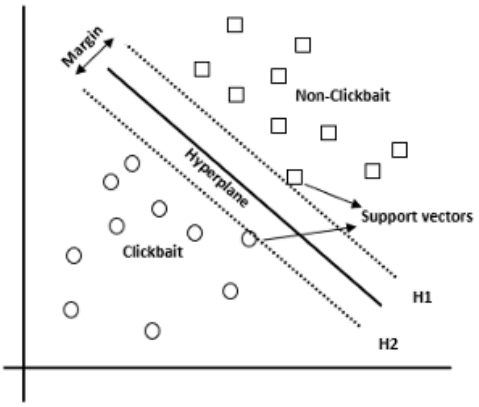
Để thực hiện huấn luyện mô hình, nhóm đã thực hiện chia tách dữ liệu. 2500 trên tổng số 3908 dữ liệu sẽ được sử dụng cho việc huấn luyện, số còn lại sẽ được chia thành hai tập test và validation.

Hai mô hình được nhóm lựa chọn là mô hình hồi quy Logistic và Support Vector Machine (SVM). Ngoài ra, nhóm cũng sẽ tiến hành huấn luyện hai mô hình này với trường hợp sử dụng hàm mất mát Gradient Descent ngẫu nhiên (SGD). Bốn mô hình sẽ thực hiện phân loại dữ liệu dựa trên từng phương pháp xử lý dữ liệu văn bản. Mục tiêu phân loại sẽ theo giá trị của cột Total.

B. Hồi quy Logistic

Hồi quy Logistic là một phương pháp thống kê được sử dụng để dự đoán xác suất xảy ra của một biến phụ thuộc nhị phân dựa trên các biến độc lập. Mô hình hồi quy Logistic sử dụng hàm logistic. Hàm logit ánh xạ y làm hàm sigmoid của x theo công thức dưới đây [10]:

$$f(x) = \frac{1}{1+e^{-x}} \tag{6}$$



Hình 2: Mô phỏng cách phân loại của mô hình SVM

Do bản chất là phân loại nhãn nhị phân, nên phương pháp này có ưu thế là mang tính đơn giản so với các phương pháp máy học khác, đồng thời dễ tiếp cận và xây dựng, linh hoạt và có tốc độ xử lý nhanh cũng như tiêu tốn ít tài nguyên thiết bị. Dữ liệu cũng được gán nhãn bằng yếu tố nhị phân (0 và 1) nên việc sử dụng phương pháp này sẽ phù hợp

C. Support Vector Machine

Support Vector Machine (SVM) là một thuật toán học máy giám sát được sử dụng để phân loại và phân tích hồi quy dữ liệu. SVM hoạt động dựa trên việc tìm ra một (hoặc một tập hợp) siêu phẳng tối ưu để phân tách dữ liệu thành các nhóm khác nhau dựa trên các đặc trưng đầu vào. Ưu thế của phương pháp bao gồm khả năng xử lý tốt dữ liệu phi tuyến nhờ vào kỹ thuật kernel, tính ổn định cao, và hiệu suất mạnh mẽ ngay cả khi số lượng đặc trưng lớn hơn số lượng mẫu [3, 10].

Ví dụ về cách thức hoạt động của SVM ta có thể thấy ở Hình 2. Mặt siêu phẳng (Hyperplane) sẽ chia và phân loại thành hai lớp: có giật gân (Clickbait) hay không giật gân (Non-Clickbait). Mỗi lớp sẽ có một cái biên (margin) và khoảng cách của hai biên với Hyperplane là bằng nhau.

Các phương trình đường Hyperplane và các đường biên $H1$, $H2$ lần lượt là:

$$\begin{aligned} H &: w^T x_i + b = 0 \\ H1 &: w^T x_i + b = 1 \\ H2 &: w^T x_i + b = -1 \end{aligned} \tag{7}$$

Đối với đề tài, SVM phù hợp trong việc xây dựng mô hình phân loại nhị phân với độ chính xác cao và khả năng kiểm soát overfitting thông qua việc điều chỉnh siêu tham số.

D. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) là một thuật toán tối ưu hóa được sử dụng rộng rãi, đặc biệt khi xử lý tập dữ liệu lớn. SGD cập nhật tham số của mô hình bằng cách tính gradient của hàm mất mát trên một hoặc vài mẫu dữ liệu tại mỗi lần lặp [10]. Những ưu điểm chính bao gồm:

- Tốc độ xử lý nhanh: Do chỉ sử dụng một hoặc vài mẫu dữ liệu trong mỗi lần cập nhật, SGD giảm đáng kể thời gian tính toán so với thuật toán Gradient Descent thông thường.
- Khả năng xử lý dữ liệu lớn: SGD có thể áp dụng trên các tập dữ liệu lớn mà không cần tải toàn bộ dữ liệu vào bộ nhớ.

- Tiêu tốn ít tài nguyên: Vì không cần lưu toàn bộ dữ liệu hay gradient, SGD tiết kiệm tài nguyên phần cứng hơn.
- Khả năng tránh cực trị cục bộ: SGD có tính ngẫu nhiên trong cập nhật, giúp tránh việc kẹt tại các điểm cực trị cục bộ trong không gian tham số.

Tuy nhiên, SGD cũng có nhược điểm là quỹ đạo hội tụ không ổn định, đòi hỏi điều chỉnh cẩn thận tốc độ học (learning rate) để đảm bảo hiệu quả.

E. Cài đặt, tinh chỉnh mô hình

Nhóm chúng em đã sử dụng các mô hình được xây dựng sẵn trong thư viện `sklearn` để thực hiện huấn luyện. Với mô hình Logistic và SVM theo SGD, nhóm đã lần lượt dùng hàm mất mát Log Loss và hàm mất mát Hinge.

Sau mỗi lần thực hiện huấn luyện mô hình, nhóm sẽ đánh giá kết quả huấn luyện phân loại bằng chỉ số Precision, Recall và Accuracy; rồi thực hiện tinh chỉnh các tham số trong mô hình để đạt kết quả tốt nhất. Khái niệm về Precision, Recall và Accuracy được chúng em ghi ở dưới đây [4]:

- Accuracy (độ chính xác) là công thức tính toán tỉ lệ dự đoán và phân loại chính xác, được tính theo công thức:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (8)$$

- Precision là công thức tính tỉ lệ dự đoán đúng các tiêu đề giật gân trong những tiêu đề được gán nhãn là giật gân, có công thức:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (9)$$

- Recall sẽ đo lường tỉ lệ của các tiêu đề giật gân được phân loại chính xác trong những tiêu đề thực sự là giật gân, theo công thức:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

Việc thực hiện huấn luyện, đánh giá, tinh chỉnh tham số và cài đặt được lặp đi lặp lại nhiều lần. Cuối cùng, nhóm đã có bảng cấu hình cài đặt tham số cho từng mô hình khi kết hợp với từng phương pháp ở Bảng II.

Bảng II: Bảng cài đặt tham số huấn luyện mô hình

Phương pháp	Mô hình	Cài đặt tham số huấn luyện
TF-IDF	Logistic	C = 1, penalty = 'l1', solver = 'liblinear'.
	Logistic-SGD	alpha = 0,01 learningrate = 'optimal'
	SVM	C = 1000, kernel = 'rbf'.
	SVM-SGD	alpha = 0,01, penalty = 'l1'. learningrate = 'optimal'
LIX/RIX	Logistic	C = 0,25, penalty = 'l1', solver = 'liblinear'.
	Logistic-SGD	alpha = 0,001 learningrate = 'optimal'
	SVM	C = 750, kernel = 'rbf'
	SVM-SGD	alpha = 0,1 learningrate = 'optimal'
BERT	Logistic	C = 0,1, solver = 'liblinear'.
	Logistic-SGD	alpha = 0,01, learningrate = 'optimal'
	SVM	C = 5, kernel = 'rbf'.
	SVM-SGD	alpha = 0,025, learningrate = 'optimal'

V. ĐÁNH GIÁ, PHÂN TÍCH KẾT QUẢ

A. Đánh giá mô hình

Kết quả phân loại tốt nhất với tập huấn luyện và tập kiểm thử được trình bày ở Bảng III và Bảng IV.

Ta thấy việc sử dụng phương pháp TF-IDF kết hợp với mô hình SVM cho ra độ chính xác trên cả tập huấn luyện và tập kiểm thử lần lượt là 0,923 và 0,794, cao hơn so với tất cả các phương pháp kết hợp còn lại. Tuy nhiên, nó lại có tiềm ẩn khả năng bị quá khớp dữ liệu (overfitting) do chênh lệch độ chính xác giữa tập huấn luyện và tập kiểm thử là khá lớn, (lớn hơn 0,1). Độ chính xác trên tập huấn luyện với từng mô hình cũng khác nhau rõ rệt, khi hai mô hình sử dụng hàm mất mát phân loại ít chính xác hơn so với hai mô hình còn lại. Xét về các chỉ số Precision và Recall, mô hình SVM khi kết hợp với TF-IDF cũng đạt giá trị cao nhất trên tập huấn luyện (lần lượt là 0,924 và 0,923) nhưng lại giảm xuống còn 0,799 và 0,794 trên tập kiểm thử.

Mặt khác, khi xử lý văn bản bằng mô hình BERT và thực hiện phân loại, ta có thể thấy được độ chính xác trên tập kiểm thử giữa các mô hình gần như là đồng đều nhau, dao động ở mức 0,78. Các chỉ số Precision và Recall của BERT cũng thể hiện tính đồng nhất trên tập kiểm thử với giá trị trong khoảng giữa 0,780 và 0,789. Tuy vậy, các giá trị chỉ số có sự chênh lệch tương đối lớn so với các giá trị trên tập huấn luyện, tương tự như khi sử dụng phương pháp TF-IDF.

Việc phân loại dựa theo chỉ số LIX và RIX có độ chính xác trên cả tập huấn luyện và kiểm thử là thấp nhất dù có sự đồng nhất và chênh lệch độ chính xác giữa hai tập là nhỏ hơn hai phương pháp còn lại. Phần lớn kết quả thu được ở cả bốn mô hình cho thấy độ chính xác sẽ rơi vào 0,76. Precision và Recall của các mô hình sử dụng phương pháp này cũng chỉ đạt tối đa 0,797 trên tập huấn luyện, giảm xuống 0,766 trên tập kiểm thử, cho thấy mức độ hiệu quả hạn chế khi so với các phương pháp khác.

Từ kết quả trên, nhóm nhận định TF-IDF kết hợp SVM là phương pháp mạnh nhất về mặt độ chính xác và chỉ số hiệu suất, trong khi BERT mang lại sự đồng nhất và đáng tin cậy hơn trong việc phân loại. Tuy vậy, các mô hình cần được cải thiện thêm, đặc biệt là mô hình sử dụng SGD, khi phần lớn các mô hình đều có độ chính xác sụt giảm khá nhiều so với mô hình không sử dụng SGD.

B. Phân tích khó khăn và giải pháp khắc phục

Trong quá trình thực hiện, nhóm gặp nhiều khó khăn trong việc thu thập tiêu đề làm dữ liệu, bao gồm việc đảm bảo tính đa dạng về nguồn, thể loại, tính đại diện, và độ cân bằng của dữ liệu. Đồng thời, việc thu thập số lượng lớn dữ liệu trong một khoảng thời gian có hạn mà vẫn đáp ứng các tiêu chí trên cũng là một thách thức không nhỏ. Quá trình này không chỉ tốn nhiều thời gian mà còn yêu cầu sự tập trung cao độ, dễ dẫn đến sự không nhất quán. Đặc biệt, việc phân loại thủ công

Bảng III: Kết quả trên tập huấn luyện

Phương pháp	Mô hình	Precision	Recall	Accuracy
TF-IDF	Logistic	0,846	0,843	0,843
	Logistic-SGD	0,769	0,764	0,764
	SVM	0,924	0,923	0,923
	SVM-SGD	0,783	0,783	0,783
LIX/RIX	Logistic	0,797	0,797	0,797
	Logistic-SGD	0,781	0,779	0,779
	SVM	0,790	0,790	0,790
	SVM-SGD	0,761	0,759	0,759
BERT	Logistic	0,875	0,874	0,874
	Logistic-SGD	0,846	0,841	0,841
	SVM	0,880	0,880	0,880
	SVM-SGD	0,848	0,848	0,848

Bảng IV: Kết quả trên tập kiểm thử

Phương pháp	Mô hình	Precision	Recall	Accuracy
TF-IDF	Logistic	0,791	0,787	0,787
	Logistic-SGD	0,743	0,737	0,737
	SVM	0,799	0,794	0,794
	SVM-SGD	0,755	0,754	0,754
LIX/RIX	Logistic	0,763	0,763	0,763
	Logistic-SGD	0,761	0,760	0,760
	SVM	0,766	0,766	0,766
	SVM-SGD	0,736	0,733	0,733
BERT	Logistic	0,782	0,781	0,781
	Logistic-SGD	0,784	0,777	0,777
	SVM	0,786	0,784	0,784
	SVM-SGD	0,789	0,788	0,788

lặp lại trên cùng một bộ dữ liệu có nguy cơ gây trung lặp giữa các lần thực hiện.

Ngoài ra, việc chỉ lấy phần tóm tắt của bài báo thay vì toàn bộ nội dung sẽ không hoàn toàn phản ánh chính xác việc tiêu đề có "giật gân" hay không, mà lại phụ thuộc nhiều vào đặc trưng ngữ nghĩa và câu từ. Chỉ số LIX và RIX nếu sử dụng trên tóm tắt cũng chưa thực sự hợp lý và khách quan trong trường hợp tóm tắt quá ngắn, trong khi kỹ thuật TF-IDF lại thì chỉ dùng được trên trường hợp đếm tần suất tiêu đề giật gân chứ không phát huy trong việc đếm độ chênh lệch giữa nội dung của phần tiêu đề với nội dung của toàn bộ bài báo.

Việc gán nhãn dữ liệu dạng nhị phân cũng bộc lộ một số hạn chế ở đề tài này, khi mà nếu chia ra theo từng tiêu chí thì số lượng lại quá ít, còn nếu xét theo tiêu chí tổng thì lại gây ra sự thiếu nhất quán về đánh giá giật gân, mặc dù tính cân bằng trong dữ liệu đã được kiểm soát.

Thêm vào đó, việc thống nhất và xây dựng bộ hướng dẫn phân loại tiêu đề là một thách thức lớn. Quá trình này đòi hỏi xác định rõ các tiêu chí phân loại và giá trị gán tương ứng thông qua nhiều cuộc thảo luận và thử nghiệm. Điều quan trọng là các tiêu chí phải được định nghĩa chi tiết, tránh mơ hồ và đảm bảo tính nhất quán giữa các thành viên. Trên thực tế, khái niệm "giật gân" đối với mỗi người sẽ có sự khác nhau. Điều này được thể hiện rõ trong khâu xây dựng bộ hướng dẫn gán nhãn (Annotation Guidelines) và thực hiện gán nhãn thử nghiệm với tập dữ liệu mẫu. Để tăng độ đồng thuận, nhóm phải giảm số lượng người gán nhãn và liên tục điều chỉnh thang đo để phản ánh các đặc trưng mới phát sinh (như đã được trình bày trong III.C).

Những khó khăn này đã biến việc hoàn thiện bộ dữ liệu thành một thử thách đòi hỏi sự phối hợp chặt chẽ và nỗ lực không ngừng từ cả nhóm.

Trong các nghiên cứu trong tương lai, nhóm sẽ xem xét đến việc viết lại rõ ràng các yếu tố để định nghĩa một tiêu đề "giật gân", xây dựng lại thang đánh giá tiêu chí tốt hơn thay vì theo thang nhị phân hiện tại. Phần tóm tắt có thể thay thế bằng nội dung của toàn bộ bài báo, để có thể phát huy hết khả năng xem xét nội dung của kỹ thuật TF-IDF.

VI. TỔNG KẾT

Như vậy, chúng em đã thực hiện phân loại mức độ giật gân của các tiêu đề bài báo điện tử cũng như trình bày quy trình xây bộ dữ liệu và gán nhãn phục vụ cho đề tài. Mặc dù cần có nhiều điểm cần bổ sung và cải thiện, đặc biệt ở quy trình gán nhãn dữ liệu, song, kết quả thu được đã đáp ứng yêu cầu của cả nhóm nói riêng và môn học nói chung. Kỹ thuật TF-IDF và mô hình ngôn ngữ BERT cũng là một công cụ đắc lực để thực hiện xử lý ngôn ngữ tự nhiên, và điều đó đã chứng minh thông qua độ chính xác cao trên tập kiểm thử. Phân loại giật gân là một đề tài tiềm năng, và nghiên cứu này sẽ tạo điều kiện cho nhóm thực hiện tiếp các nghiên cứu khác trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Yên Thuỷ, "Những bất cập phổ biến của trang tin điện tử tổng hợp, mạng xã hội," Vietnam+ (VietnamPlus).

[2] P. Biyani, K. Tsioutsiouliklis, and J. Blackmer, "'8 Amazing Secrets for Getting More Clicks': Detecting Clickbaits in News Streams Using Article Informality," *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, Art. no. 1, Feb. 2016, doi: 10.1609/aaai.v30i1.9966.

[3] S. R. Dam, S. P. Panday, and T. B. Thapa, "Detecting Clickbaits on Nepali News using SVM and RF," vol. 9, 2021.

[4] M. Bronakowski, M. Al-khassaweneh, and A. Al Bataineh, "Automatic Detection of Clickbait Headlines Using Semantic Analysis and Machine Learning Techniques," *Appl. Sci.*, vol. 13, no. 4, p. 2456, Feb. 2023, doi: 10.3390/app13042456.

[5] H. A. Ahmadi and A. Chowanda, "Clickbait Classification Model on Online News with Semantic Similarity Calculation Between News Title and Content," *Build. Inform. Technol. Sci. BITS*, vol. 4, no. 4, Art. no. 4, Mar. 2023, doi: 10.47065/bits.v4i4.3030.

[6] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, vol. 28, no. 1, pp. 11–21, Jan. 1972, doi: 10.1108/eb026526.

[7] J. Anderson, "Lix and Rix: Variations on a Little-known Readability Index," *J. Read.*, vol. 26, no. 6, pp. 490–496, 1983.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[9] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[10] Vũ Hữu Tiệp, *Machine Learning cơ bản*. 2018.