Predicting Popularity of VNExpress Articles

Vũ Đình Nhật [1,2,3], Hồ Huỳnh Thư Nhi [1,2,4], Lê Diễm Quỳnh Như [1,2,5]
TS. Nguyễn Gia Tuấn Anh[1,2], CN. Trần Quốc Khánh[1,2]
[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City, Vietnam*
[3] 23521104@gm.uit.edu.vn
[4] 23521107@gm.uit.edu.vn
[5] 23521122@gm.uit.edu.vn
https://github.com/nhatvu205/Predicting-Popularity-of-Vietnamese-Articles

1. Tóm tắt

Trong đồ án này, mục tiêu của nhóm chúng em là xây dựng hệ thống dự đoán độ phổ biến của bài viết trên VNExpress, dựa trên các chỉ số như bình luận và lượt tương tác trong bình luận. Dữ liệu được thu thập thông qua Web Scraping từ VNExpress, sau đó thực hiện tiền xử lý bao gồm lọc dữ liệu, chuẩn hóa văn bản và trích xuất đặc trưng. Các mô hình học máy như Hồi quy Ridge, Random Forest, XGBoost sẽ được xây dựng và đánh giá. Kết quả đánh giá hiệu suất mô hình sẽ định hướng các bước phát triển tiếp theo, như tối ưu hóa đặc trưng hoặc thử nghiệm mô hình nâng cao.

Từ khóa: Dự đoán độ phổ biến, Học máy, VnExpress, Xử lý ngôn ngữ tự nhiên, Hồi quy, Tiếng Việt.

2. Giới thiệu

Trong bối cảnh báo chí trực tuyến ngày càng phát triển, việc dự đoán độ phổ biến của các bài viết trên các nền tảng trực tuyến có ý nghĩa lớn trong việc tối ưu hóa nội dung và tăng cường tương tác với độc giả. Bài toán đặt ra nhằm xây dựng một hệ thống học máy dự đoán các chỉ số tương tác của bài viết tiếng Việt, cụ thể với trang báo VnExpress sẽ là số bình luận và số lượt tương tác trong phần bình luận. VnExpress được nhóm em chọn làm nguồn dữ liệu do đây là một trong những trang báo điện tử hàng đầu Việt Nam, cung cấp lượng dữ liệu phong phú, đa dạng về chủ đề và có số lượng tương tác lớn từ người đọc. Mục tiêu là tạo ra mô hình có khả năng dự đoán chính xác mức độ phổ biến, hỗ trợ biên tập viên, cùng với đó là những người quản lý và sáng tạo nội dung đưa ra quyết định mang tính chiến lược hơn. Thông tin đầu vào bao gồm các đặc trưng trích xuất từ bài viết như tiêu đề, nội dung, danh mục, thời gian đăng,... thu thập qua Web Scraping. Kết quả mong muốn sẽ là các giá trị dự đoán dựa trên số bình luận, và số lượt tương tác trong phần bình luận.

3. Các công trình nghiên cứu liên quan

3.1. Các nghiên cứu về dự đoán độ phổ biến nội dung

Dự đoán mức độ phổ biến của bài báo là một nhánh quan trọng trong khai phá dữ liệu văn bản. Nhiều nghiên cứu đã áp dụng các mô hình học máy có giám sát để dự đoán lượt xem, chia sẻ hoặc tương tác của bài viết. Trong [1], tác giả sử dụng các đặc trưng đơn giản như tiêu đề, độ dài và thời gian đăng để huấn luyện mô hình hồi quy. Tương tự, [2] khai thác metadata và nội dung văn bản để đưa ra dự đoán chính xác về độ phổ biến tin tức.

3.2. Các kỹ thuật xử lý văn bản

Xử lý ngôn ngữ tự nhiên đóng vai trò then chốt trong việc trích xuất đặc trưng từ bài viết. Các kỹ thuật phổ biến như tách từ, loại bỏ từ dừng, TF-IDF hay Word2Vec đã được sử dụng rộng rãi [3], [4]. Ngoài ra, các mô hình embedding hiện đại như BERT cũng được tích hợp trong các hệ thống dự đoán, như trong nghiên cứu [5] với mô hình Multi-LSTM kết hợp BERT.

3.3. Định hướng nghiên cứu

Hầu hết các nghiên cứu hiện nay chủ yếu tập trung vào ngữ liệu tiếng Anh và dữ liệu từ các mạng xã hội hoặc nền tảng truyền thông quốc tế như Twitter, Facebook, hoặc các báo điện tử lớn như New York Times, BBC, CNN. Trong khi đó, việc nghiên cứu và áp dụng các phương pháp tương tự cho tiếng Việt còn khá hạn chế.

Đề tài này hướng đến việc mở rộng ứng dụng các kỹ thuật đã được kiểm chứng trong các nghiên cứu trước vào ngữ liệu tiếng Việt, cụ thể là tập dữ liệu các bài báo từ trang VNExpress. Bằng cách khai thác nội dung bài viết, tiêu đề và thông tin metadata, kết hợp với các mô hình học máy, đề tài kỳ vọng xây dựng được một hệ thống dự đoán độ phổ biến bài báo một cách hiệu quả và phù hợp với ngữ cảnh tiếng Việt.

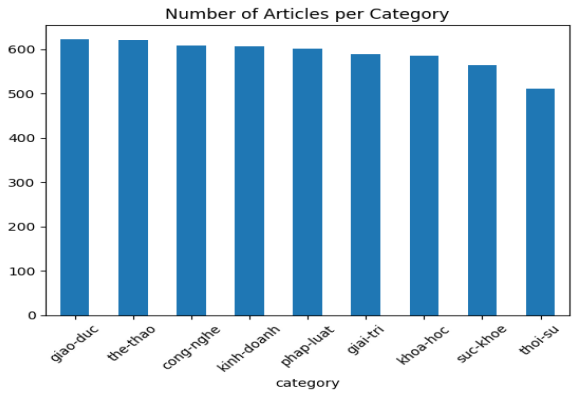
4. Bộ dữ liệu

4.1. Thu thập dữ liệu

4.1.1. Nguồn dữ liệu

Dữ liệu được thu thập từ trang thông tin điện tử VnExpress (địa chỉ: https://vnexpress.net) - một trong những tờ báo điện tử lớn nhất Việt Nam. Tổng cộng có 9 chuyên mục được khai thác bao gồm: Khoa học, Sức khỏe, Giáo dục, Thời sự, Giải trí, Thể thao, Công nghệ, Pháp luật và Kinh doanh. Việc lựa chọn nhiều chuyên mục nhằm đảm bảo tính đa dạng và khách quan của tập dữ liệu.

4.1.2. Công cụ thu thập



Hình 1. Số lượng bài báo mỗi chuyên mục

Việc thu thập dữ liệu được thực hiện bằng kỹ thuật Web Scraping kết hợp giữa thư viện Selenium (giả lập trình duyệt để tương tác với trang web) và BeautifulSoup (trích xuất nội dung HTML). Quy trình bao gồm các bước: truy cập trang chủ, duyệt danh sách bài viết theo chuyên mục, trích xuất thông tin, sau đó chuẩn hóa dữ liệu và lưu vào dạng CSV.

4.1.3. Nhãn của tập dữ liệu

Tập dữ liệu được gán nhãn dựa trên thuộc tính popularity_score (điểm phổ biến), được xây dựng dựa trên các thông số tương tác với người dùng là số bình luận và tổng số tương tác bình luận. Các giá trị của popularity_score mang tính liên tục, đáp ứng bản chất của bài toán hồi quy.

4.1.4. Tổng quan dữ liệu

Dữ liệu được thu thập từ giữa tháng 12/2024 đến ngày 18/04/2025.

Dữ liệu gồm 5310 dòng và 9 thuộc tính.

Thuộc tính	Mô tả
title	Tiêu đề bài báo
date	Thời gian đăng bài
wordcount	Số lượng từ trong bài báo
comments	Số bình luận
interactions	Tổng số tương tác bình luận
images	Số ảnh của bài báo
videos	Số video của bài báo
tags	Liệt kê các tags của bài báo
category	Chuyên mục

Bảng 1. Các thuộc tính được thu thập

4.2. Tiền xử lý dữ liệu

4.2.1. Gán nhãn cho dữ liệu

Việc gán nhãn được thực hiện trên các bài viết bằng cách tính toán thuộc tính popularity_score theo công thức:

$$popularity\_score = (0.1 * comments + 0.1 * interactions) / (time\_since\_posted + 1e\text{-}10)$$

Giá trị này được gán vào mỗi bài báo như một nhãn để huấn luyện mô hình học máy.

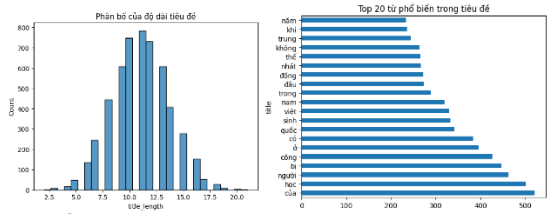
Công thức được xây dựng dựa trên nguyên lý rằng bài viết càng có nhiều tương tác (bình luận, lượt thích/chia sẻ) thì càng có khả năng phổ biến. Thành phần thời gian được đưa vào mẫu số nhằm giảm điểm phổ biến cho những bài viết đã đăng từ lâu. Để tránh sự chênh lệch quá lớn giữa bài mới và bài cũ, giá trị thời gian đã được biến đổi logarit (log-transform), giúp làm trơn ảnh hưởng của thời gian. Ngoài ra, hệ số 0.1 giúp chuẩn hóa các giá trị tương tác về cùng thang đo, và số nhỏ $1e{-}10$ được thêm để tránh chia cho 0.

Cách xây dựng công thức này được lấy cảm hứng từ các thuật toán đánh giá độ phổ biến trong thực tế như Reddit Ranking [8] hay Twitter Virality Score [9]. Các phương pháp này đều nhấn mạnh sự kết hợp giữa mức độ tương tác của người dùng và yếu tố thời gian đăng tải, nhằm phản ánh chính xác khả năng lan truyền và mức độ thu hút của nội dung trên nền tảng số.

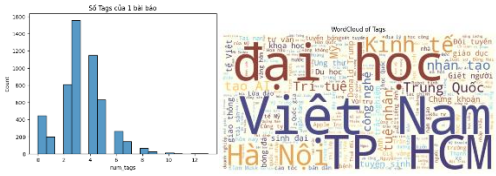
4.2.2. Xử lý dữ liệu

- Làm sạch dữ liệu: Loại bỏ các bản ghi không hợp lệ như tiêu đề rỗng, nội dung vô nghĩa, wordcount = 0, ngày đăng sai định dạng, hoặc trùng lặp tiêu đề. Đồng thời, xử lý các dòng bị thiếu giá trị – do tỷ lệ thiếu nhỏ và phần lớn là giá trị không rõ ràng (ví dụ: ghi "Không rõ", để trống), nên việc loại bỏ giúp tránh gây nhiễu và đảm bảo mô hình không học sai bản chất dữ liệu. Kiểm tra và đảm bảo các cột số (wordcount, comments, interactions, images, videos) là giá trị số không âm.
- Chuẩn hóa văn bản: Tiêu đề và nội dung được chuẩn hóa unicode, loại bỏ ký tự đặc biệt, tách từ tiếng Việt bằng Pyvi/Vn CoreNLP, và loại bỏ stopwords để hỗ trợ trích xuất đặc trưng ngôn ngữ.
- Xử lý dữ liệu số: Phân tích và xử lý giá trị ngoại lai (outlier) bằng phương pháp IQR, áp dụng clipping hoặc loại bỏ. Với các cột có phân phối lệch (skewness > 1 hoặc < -1), áp dụng biến đổi (log, Box-Cox) để đưa dữ liệu về phân phối chuẩn hơn.

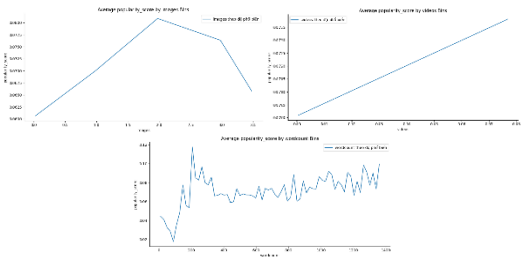
4.2.3. Khám phá dữ liệu (EDA)



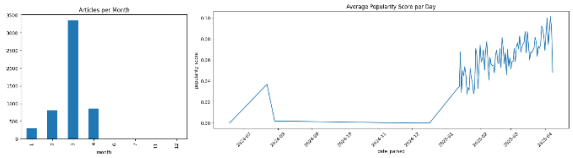
Tiêu đề có độ dài hợp lý (5–18 từ) thường thu hút hơn, trong khi các tiêu đề quá ngắn hoặc quá dài ít phổ biến và kém hiệu quả. Các từ xuất hiện nhiều trong tiêu đề chủ yếu là từ chung chung nên ít mang tính phân biệt. Do đó, có thể trích xuất các đặc trưng như độ dài tiêu đề, cảm xúc, câu hỏi, điểm ngữ nghĩa và loại bỏ từ dừng để nâng cao khả năng dự đoán độ phổ biến.



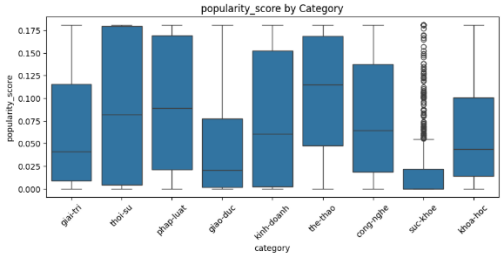
Tags phản ánh chủ đề nội dung bài viết và có liên hệ chặt chẽ đến độ phổ biến – bài viết có nhiều tags thường được tìm kiếm và gợi ý nhiều hơn. Số lượng tags phổ biến là khoảng 3, và các tag thường gặp xoay quanh giáo dục, công nghệ, địa danh lớn. Do đó, nên giữ num_tags làm đặc trưng đầu vào và chuyển nội dung tags thành feature bằng các kỹ thuật như TF-IDF, one-hot hoặc word embedding.



Các thuộc tính wordcount, images và videos đều cho thấy mối liên hệ nhất định với độ phổ biến của bài viết và mang giá trị thông tin rõ ràng. Wordcount có ảnh hưởng phi tuyến, với độ dài tối ưu nằm trong khoảng 100–400 từ; images cũng cho thấy hiệu quả cao nhất khi có 1–2 ảnh; còn videos ảnh hưởng nhẹ và chủ yếu phân biệt có/không có video. Do đó, cả ba thuộc tính nên được giữ nguyên và chỉ cần chuẩn hóa nếu cần thiết khi đưa vào mô hình.



Thời gian đăng bài cho thấy ảnh hưởng rõ rệt đến độ phổ biến, đặc biệt vào các giai đoạn cao điểm như tháng 3. Mức độ phổ biến dao động mạnh theo ngày, gợi ý vai trò quan trọng của yếu tố thời gian cụ thể như ngày trong tuần hoặc giờ đăng. Việc trích xuất các đặc trưng như is_weekday và posted_hour là hợp lý để giúp mô hình học được các quy luật thời gian tiềm ẩn.



Dữ liệu cho thấy chuyên mục bài viết (category) có ảnh hưởng đáng kể đến độ phổ biến, với các chuyên mục như thời sự, pháp luật, thể thao thường đạt điểm cao hơn so với sức khỏe hay giáo dục. Tuy nhiên, do thuộc tính này không được sử dụng trong quá trình huấn luyện mô hình, việc phân tích chỉ mang tính tham khảo để hiểu rõ hơn về xu hướng dữ liệu và không ảnh hưởng đến việc xây dựng đặc trưng đầu vào.

4.2.4. Trích xuất đặc trưng

4.2.4.1. Đặc trưng từ tiêu đề (title)

Tiêu đề bài viết chứa đựng nhiều thông tin quan trọng về nội dung, mức độ thu hút và cảm xúc. Do đó, nhóm thực hiện trích xuất một tập hợp các đặc trưng từ trường title nhằm phục vụ cho quá trình huấn luyện mô hình dự đoán. Cụ thể gồm các đặc trưng sau:

- title_length: Số lượng từ trong tiêu đề.
- has_number: Biến nhị phân, kiểm tra tiêu đề có chứa chữ số hay không.
- has_emotion: Biến nhị phân, kiểm tra tiêu đề có chứa từ khóa biểu cảm mạnh hay không.
- is_question: Biến nhị phân, kiểm tra tiêu đề có mang tính nghi vấn hay không.
- sentiment_score: Điểm cảm xúc của tiêu đề, phản ánh xu hướng cảm xúc của tiêu đề:
  - Positive: Nghiêng về cảm xúc tích cực.
  - Negative: Nghiêng về cảm xúc tiêu cực.
  - Gần 0: Trung tính.

Quy trình chi tiết:

1. Xử lý văn bản: Tiêu đề được chuẩn hóa (chuyển thành chữ thường, xử lý khoảng trắng và ký tự đặc biệt).
2. Phân tích cảm xúc (sentiment_score): Tiêu đề được đưa vào mô hình phân tích cảm xúc (có sử dụng word_tokenize để tách từ), và tính toán xác suất ba lớp: positive, negative, neutral. Từ đó, tính điểm cảm xúc bằng công thức:

$$sentiment\_score = P(positive) - P(negative)$$

3. Huấn luyện mô hình tính title_score:

- Dữ liệu được chia theo từng chuyên mục (category).
- Với mỗi chuyên mục đủ dữ liệu (≥ 5 bài viết), huấn luyện mô hình hồi quy tuyến tính với các đặc trưng title_length, has_number, has_emotion, is_question, sentiment_score.
- Dự đoán và lưu kết quả vào cột title_score, đại diện cho chất lượng dự đoán của tiêu đề so với điểm phổ biến thực tế.

Việc bổ sung sentiment_score giúp mô hình hiểu được tác động tiềm ẩn của cảm xúc trong tiêu đề đến mức độ phổ biến, đặc biệt hiệu quả trong các chuyên mục như giải trí, thời sự, xã hội, nơi ngôn ngữ cảm xúc được sử dụng thường xuyên.

4.2.4.2. Đặc trưng từ thời gian (date)

Từ trường date (thời điểm đăng bài viết), nhóm thực hiện rút trích các đặc trưng thời gian có tính chu kỳ và ảnh hưởng đáng kể đến độ phổ biến của bài viết. Các đặc trưng này giúp mô hình nhận diện được các quy luật về thời điểm đăng bài ảnh hưởng đến lượng tương tác và mức độ thu hút độc giả.

Các đặc trưng được trích xuất bao gồm:

- is_weekday: Biến nhị phân biểu diễn bài viết được đăng vào ngày trong tuần (thứ 2 đến thứ 6) hay cuối tuần (thứ 7, chủ nhật). Yếu tố này giúp mô hình nhận biết được sự khác biệt về thói quen đọc tin của độc giả vào các ngày làm việc và ngày nghỉ.
- posted_hour: Biến số nguyên thể hiện giờ trong ngày khi bài viết được đăng (mã hóa nhãn từ 0 đến 23). Thông tin này phản ánh thói quen truy cập và tương tác của người đọc theo khung giờ khác nhau trong ngày.

Quy trình trích xuất đặc trưng:

1. Chuẩn hóa định dạng trường thời gian, đảm bảo tính đồng nhất.
2. Tách riêng ngày trong tuần và giờ đăng bài từ trường date.
- Mã hóa biến is_weekday thành dạng nhị phân (0 – cuối tuần, 1 – ngày thường).
- Mã hóa posted_hour dưới dạng số nguyên từ 0 đến 23, tương ứng với các giờ trong ngày.

4.2.4.3. Đặc trưng từ thẻ bài viết (tags)

Tags là tập hợp các từ khóa phản ánh nội dung chính của bài viết, thường liên quan đến nhân vật, địa điểm, sự kiện hoặc tổ chức. Nhằm khai thác giá trị của tags trong việc dự đoán mức độ phổ biến, nhóm tiến hành trích xuất các đặc trưng sau:

- has_person: Biến nhị phân, giá trị 1 nếu bài viết có tag thuộc loại person (cá nhân, nhân vật).
- has_place: Biến nhị phân, giá trị 1 nếu xuất hiện tag thuộc loại place (địa điểm, khu vực).
- has_event: Biến nhị phân, giá trị 1 nếu có tag thuộc loại event (sự kiện).
- has_org: Biến nhị phân, giá trị 1 nếu có tag thuộc loại org (tổ chức, công ty).

Quy trình trích xuất:

1. Tiền xử lý và lọc tag phổ biến:

- Tập hợp tất cả tags từ dữ liệu, chuẩn hóa về chữ thường, loại bỏ ký tự đặc biệt và khoảng trắng thừa. Chỉ giữ lại những tags xuất hiện từ 5 lần trở lên để đảm bảo tính đại diện và giảm nhiễu.

2. Chuẩn hóa và phân loại tag:

- Sử dụng từ điển thủ công để ánh xạ tag vào các nhóm: person, place, event, org, unknown.
- Đồng thời kết hợp công cụ Underthesea để hỗ trợ gán nhãn tự động cho các tags chưa có trong từ điển, tăng độ bao phủ và tính linh hoạt.

3. Tính toán đặc trưng:

- Với mỗi bài viết, xác định xem có ít nhất một tag thuộc từng loại (person, place, event, org) để gán giá trị tương ứng cho các biến nhị phân has_*.
- Ngoài ra, các tags đã lọc còn được sử dụng để tính TF-IDF vector làm đặc trưng đầu vào bổ sung, giúp mô hình khai thác sâu hơn về nội dung ngữ nghĩa của bài viết.

Việc kết hợp giữa tag từ điển và gán nhãn tự động giúp mô hình tận dụng hiệu quả thông tin ngữ nghĩa từ tags, đặc biệt trong các chuyên mục có yếu tố thực thể rõ rệt như thời sự, thể thao, giải trí.

4.2.4.4. Đặc trưng số học có sẵn

Ngoài các đặc trưng trích xuất từ văn bản, nhóm cũng sử dụng trực tiếp các trường định lượng sẵn có trong dữ liệu gồm:

- wordcount: Tổng số từ trong nội dung bài viết.
- images: Số lượng hình ảnh được chèn trong bài.
- videos: Số lượng video được đính kèm.

Ba đặc trưng này được chuẩn hóa (standardization) trước khi đưa vào mô hình học máy nhưng không cần trích xuất thêm vì đã có sẵn từ dữ liệu thô.

4.3. Đánh giá dữ liệu

Nhóm chúng em đánh giá bộ dữ liệu thu thập được dựa trên 6 tiêu chí từ quyển sách “"Data Mining – Concept and Techniques 3rd edition" bởi J. Han, J. Pei và M. Kamber [6]. Trong đó, 6 tiêu chí bao gồm:

- **Sự chính xác (Accuracy):** Dữ liệu được thu thập và đối chiếu thủ công bởi nhóm chúng em để đảm bảo không có sai sót về thời gian, số lượng
- **Sự đầy đủ (Completeness):** Dữ liệu về bài báo được thu thập toàn diện trên các chuyên mục, trải dài từ Thời sự, Giáo dục, Kinh tế cho đến Thể thao, Giải trí.
- **Sự nhất quán (Consistency):** Toàn bộ quá trình thu thập và dữ liệu cuối cùng đều có chung hình thức, đầy đủ các đặc trưng được thống nhất ngay từ khi bắt đầu thu thập.
- **Sự kịp thời (Timeliness):** Dữ liệu chỉ thu thập được ở khoảng thời gian giữa 12/2024 – 4/2025, chưa trải đều trong năm và đa dạng sự kiện đặc biệt.
- **Sự tin cậy (Believability):** Nguồn dữ liệu đến từ trang báo VNExpress, một trong những trang báo điện tử lớn nhất Việt Nam.
- **Sự dễ hiểu (Interpretability):** Các đặc trưng dữ liệu được diễn giải rõ ràng qua codebook, từ “Tiêu đề”, “Ngày đăng” đến “Tags”, “Chuyên mục” đều là những cụm từ quen thuộc với người đọc.

4.4. Phân chia dữ liệu

Nhóm chúng em tiến hành phân chia dữ liệu tập train:dev:test theo tỉ lệ 8:1:1, sau đó thực hiện GridSearch với k-fold = 5 để tìm tham số tốt nhất của mỗi mô hình trên tập train và dev

4.5. Lựa chọn đặc trưng (Feature Selection)

Để đảm bảo mô hình có thể học được các mô thức tốt nhất từ đặc trưng, nhóm chúng em tiến hành bước lựa chọn đặc trưng với 02 phương pháp chính: Feature Importance và Mutual Information.

Khi thực hiện Mutual Information, các đặc trưng nhị phân liên quan tương tác (thành phần cấu thành biến mục tiêu) có điểm cao như has_comments, has_interactions, còn một vài đặc trưng của title và date thì kém nổi bật hơn với MI score = 0.

Khi sử dụng Feature Importance, một vài dặc trưng được vetor hóa TF-IDF từ tags có điểm quan trọng nổi bật hơn các đặc trưng khác. Sau đó, chúng em loại bỏ các đặc trưng rơi vào nhóm Feature Importance score = 0 để mô hình được học tối ưu hơn.

5. Các mô hình học máy áp dụng

Bài toán đặt ra trong dự án là dự đoán điểm phổ biến của bài bo – một đại lượng liên tục được tính toán dựa trên số lượng bình luận và tương tác. Do đó, đây là một bài toán thuộc loại hồi quy. Việc sử dụng các mô hình hồi quy giúp xây dựng mối quan hệ hàm số giữa tập các đặc trưng đầu vào (số từ, số ảnh, số video, chuyên mục, tags, thời gian đăng bài, v.v.) và giá trị mục tiêu.

5.1. Mô hình hồi quy Ridge (Ridge Regression)

Hồi quy Ridge là một biến thể của hồi quy tuyến tính, bổ sung thành phần chính quy hóa L2 nhằm hạn chế hiện tượng quá khớp khi có nhiều đặc trưng hoặc các đặc trưng có tương quan cao. Phương trình nghiệm được điều chỉnh để phạt (penalize) các trọng số lớn, giúp mô hình ổn định hơn với dữ liệu có đa tuyến tính.

Lý do lựa chọn mô hình cho bài toán:

- Phù hợp khi các đặc trưng đầu vào (như số từ, số ảnh, ngày đăng bài) có thể có quan hệ tuyến tính với mục tiêu.
- Thành phần chuẩn hóa (regularization) giúp ổn định mô hình khi có các đặc trưng tương quan như "số bình luận" và "số tương tác".
- Có thể mở rộng dễ dàng với các đặc trưng phi tuyến nhờ biến đổi polynomial và spline.

5.2. Mô hình hồi quy Random Forest (Random Forest Regression)

Random Forest là mô hình tổ hợp của nhiều cây quyết định. Mỗi cây được huấn luyện trên một tập con dữ liệu và đặc trưng ngẫu nhiên. Kết quả cuối cùng là trung bình các dự đoán của từng cây.

Lý do lựa chọn mô hình cho bài toán:

- Có khả năng mô hình hóa mối quan hệ phi tuyến và tương tác giữa các đặc trưng (ví dụ: số video và chuyên mục).
- Không yêu cầu chuẩn hóa đặc trưng đầu vào.
- Tự động đánh giá độ quan trọng của các đặc trưng, giúp hiểu rõ yếu tố nào ảnh hưởng đến độ phổ biến của bài báo.

5.3. Mô hình hồi quy XGBoost (Extreme Gradient Boost Regression)

XGBoost là mô hình boosting, xây dựng dần từng cây quyết định nhỏ để học phần dư sai số (residuals) từ mô hình trước đó.

Lý do lựa chọn mô hình cho bài toán:

- Hiệu quả cao với dữ liệu tabular như metadata bài báo.
- Có khả năng xử lý tốt mối quan hệ phi tuyến và tương tác phức tạp giữa nhiều đặc trưng.
- Hỗ trợ regularization mạnh (L1 và L2), giúp kiểm soát overfitting tốt hơn.

6. Kết quả thí nghiệm

Sau khi huấn luyện các mô hình học máy bao gồm Ridge Regression, Random Forest Regressor và XGBoost Regressor, kết quả được đánh giá trên tập phát triển (dev set) và tập kiểm tra (test set) bằng các chỉ số RMSE, R^2. Đặc biệt, mô hình Random Forest cũng được kiểm định thống kê phần dư để đánh giá chất lượng dự đoán một cách sâu sắc hơn ở bên dưới.

Model	RMSE (Dev)	R^2 (Dev)	RMSE (Test)	R^2 (Test)
Ridge Regression	0.0007	0.4348	0.0006	0.4342
Random Forest	0.0007	0.4282	0.0006	**0.4610**
XGBoost	0.0007	**0.4348**	0.0006	0.4468

Bảng 2. Kết quả đánh giá mô hình trên tập dữ liệu dev và test

Mô hình Random Forest cho kết quả tốt nhất với RMSE nhỏ nhất và hệ số R^2 cao nhất trên tập test. Điều này cho thấy mô hình này có khả năng học tốt mối quan hệ giữa các đặc trưng metadata và độ phổ biến bài báo.

Random Forest là mô hình tổ hợp nhiều cây quyết định, có khả năng học tốt các mối quan hệ phi tuyến và giảm thiểu hiện tượng overfitting nhờ việc lấy trung bình kết quả của nhiều cây.

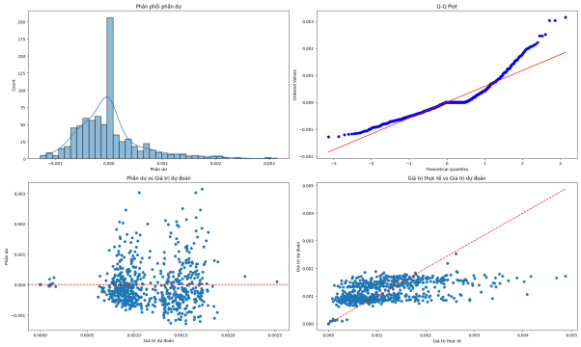
Khi so sánh với XGBoost, Random Forest ít nhạy cảm với việc tuning siêu tham số và có thể tổng quát hóa tốt hơn khi dữ liệu không quá phức tạp như của nhóm. Khi so sánh với Ridge Regression, đây là một mô hình hồi quy tuyến tính có thêm thành phần phạt (regularization), nhưng bản chất của nó vẫn chỉ mô hình hóa mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu nên không mạnh mẽ bằng các mô hình phi tuyến trong khả năng học và khai thác tốt các mối quan hệ phức tạp, tương tác giữa các đặc trưng, cũng như các mẫu dữ liệu có tính phi tuyến hoặc phân phối không đều.

Trên tập dev, XGBoost có R^2 cao hơn Random Forest, nhưng trên tập test thì Random Forest lại vượt trội hơn, cho rằng XGBoost dễ bị overfitting nhẹ trên tập dev nếu siêu tham số chưa tối ưu hoàn toàn

6.1. Kiểm định thống kê trên kết quả mô hình Random Forest

Trong các bài toán hồi quy, việc kiểm tra các giả định kinh điển về phần dư là rất quan trọng để đánh giá mức độ phù hợp và độ tin cậy của mô hình. Các kiểm định thống kê như Shapiro-Wilk, Jarque-Bera (kiểm định tính chuẩn), Durbin-Watson, Ljung-Box (kiểm định tính độc lập), Breusch-Pagan, Goldfeld-Quandt (kiểm định tính đồng nhất phương sai) đều là những phương pháp phổ biến, được đề xuất trong các tài liệu kinh điển về phân tích hồi quy [7].

Hình 2 và bảng 3 là kết quả khi chúng em thực hiện các bài test kiểm định thống kê trên tập dữ liệu test và dữ liệu dự đoán của mô hình tốt nhất - Random Forest.



Hình 2. Lần lượt từ trái sang phải, trên xuống dưới: Biểu đồ phân phối phần dư, giá trị Q-Q plot, phần dư so với dự đoán, thực tế so với dự đoán

Kiểm định	Thống kê	p-value	Kết luận	Ghi chú
Shapiro-Wilk	0.8912	0.0000	Không chuẩn	Phần dư không tuân theo phân phối chuẩn
Jarque-Bera	685.3640	0.0000	Không chuẩn	Phần dư không tuân theo phân phối chuẩn
Durbin-Watson	2.0413	–	Độc lập	Không có tự tương quan bậc 1
Ljung-Box (lag=5)	3.5785	0.6115	Độc lập	Không có tự tương quan
Ljung-Box (lag=10)	5.8027	0.8316	Độc lập	Không có tự tương quan
Ljung-Box (lag=15)	17.0509	0.3158	Độc lập	Không có tự tương quan
Breusch-Pagan	378.5979	0.0000	Đồng nhất	Phương sai đồng nhất
Goldfeld-Quandt	0.8400	0.9494	Đồng nhất	Phương sai đồng nhất
Đa cộng tuyến (VIF)	–	–	Có (một số biến)	Một số biến có VIF vô cực (đa cộng tuyến mạnh)
Cook's Distance	–	–	Có ngoại lai	30 điểm có Cook's D > 4/n
F-test	1.4101	0.0224	Có ý nghĩa	Mô hình có ý nghĩa thống kê

Bảng 3. Kết quả kiểm định thống kê trên mô hình tốt nhất (Random Forest)

Kết quả kiểm định cho thấy:

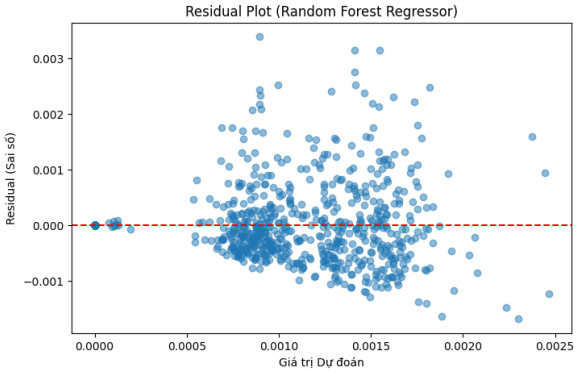
- Phần dư của mô hình Random Forest không tuân theo phân phối chuẩn (p-value của kiểm định Shapiro-Wilk và Jarque-Bera đều rất nhỏ).

- Kiểm định Durbin-Watson cho giá trị 2.0413, gần với giá trị lý tưởng 2, cho thấy phần dư độc lập.
- Kiểm định Durbin-Watson và Ljung-Box đều cho thấy phần dư độc lập và không có tự tương quan (p-value > 0.05).
- Kiểm định Goldfeld-Quandt cho p-value = 0.9494 cho thấy phương sai đồng nhất.
- Đa cộng tuyến: Một số biến có giá trị VIF rất lớn, cho thấy có dấu hiệu đa cộng tuyến mạnh ở các biến được vector hóa TF-IDF, còn lại các biến truyền thống không đáng quan ngại.
- Ngoài ra còn có các kiểm định khác như kiểm F cho thấy mô hình có ý nghĩa thống kê, kiểm tra điểm ngoại lai bằng Cook's Distance cho thấy có 30 điểm cần xem xét.

Nhìn chung, mặc dù một số giả định kinh điển của hồi quy tuyến tính không được thỏa mãn hoàn toàn, mô hình Random Forest vẫn đảm bảo được tính độc lập của phần dư và cho kết quả dự báo tốt trên tập dữ liệu của chúng em.

6.2. Phân tích lỗi

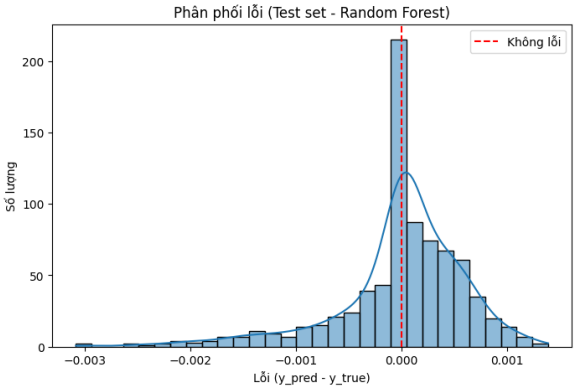
Nhóm chúng em thực hiện phân tích lỗi dựa trên tập y_pred của mô hình tốt nhất - Random Forest. Trong đó, pân bố residual (phần dư, được tính bằng y_pred - y_true) khá ngẫu nhiên quanh trục 0 cho thấy đây là dấu hiệu tốt. Tuy nhiên, cụm dữ liệu rời rạc cho thấy đầu ra của mô hình bị rời rạc hóa (do đặc trưng của Random Forest – không nội suy mượt mà). Có nhiều điểm nằm xa trục 0 nên có thể overfit nhẹ. Vì vậy, mô hình tốt hơn Ridge và XGBoost về mặt nắm bắt phi tuyến, nhưng có thể cải thiện bằng thêm dữ liệu hoặc đặc trưng.



Hình 3. Biểu đồ thể hiện giá trị phần dư (y_pred - y_true) so với giá trị thực tế (y_true)

Khi phân tích chi tiết kết hợp phân phối lỗi trong hình 9, ta thấy:

- Các điểm bên phải (khi y_true > 0.002) đa phần có sai số âm ⇒ y_pred < y_true. Điều này cho thấy mô hình under-predicts (dự đoán thấp hơn) với các giá trị lớn của nhãn.
- Phân phối sai số không đối xứng hoàn toàn. Dựa vào hình 3, phần lớn sai số là dương ở vùng y_true thấp, và âm ở vùng y_true cao ⇒ mô hình có vẻ chưa học tốt xu hướng tăng đều của nhãn theo đặc trưng.



Hình 4. Biểu đồ phân phối giá trị lỗi (y_pred - y_true)

7. Kết luận và hướng phát triển

7.1. Kết luận

Trong nghiên cứu này, chúng em đã xây dựng mô hình hồi quy dự đoán độ phổ biến của bài báo tiếng Việt dựa trên các đặc trưng metadata như tiêu đề, thời gian đăng, số lượng từ, số ảnh, số video, số tương tác, bình luận và thông tin ngữ nghĩa (title, tags, category). Kết quả thực nghiệm cho thấy mô hình XGBoost có hiệu quả vượt trội với R^2 đạt 46,28% và RMSE nhỏ nhất trên tập kiểm tra.

Vì lý do thời gian hạn chế cộng với việc chưa có nhiều kinh nghiệm trong việc xử lý ngôn ngữ tự nhiên, nhóm chúng em nhận thấy chưa tận dụng tốt được các đặc trưng như tags để có thể cải thiện độ đo của bài báo.

7.2. Hướng phát triển

Mặc dù các chỉ số đánh giá cho thấy khả năng dự đoán đang ở mức trung bình, nhưng để nâng cao độ chính xác hơn nữa, các hướng phát triển tiếp theo có thể bao gồm:

- Sử dụng phương pháp multimodal (kết hợp xử lý ngôn ngữ tự nhiên và hình ảnh) để tăng đặc trưng cho mô hình huấn luyện
- Nhận diện thực thể trong văn bản của các đặc trưng tags có thể được cải thiện bằng VNcoreNLP, phoBERT
- Ứng dụng các kỹ thuật giải thích mô hình (SHAP, permutation importance) để hiểu rõ hơn ảnh hưởng của từng đặc trưng đầu vào.

- Xây dựng quy trình tự động thu thập dữ liệu trên thời gian thực bằng N8N để thiết lập các luồng tự động để định kỳ truy cập vào trang báo và trích xuất thông tin.

8. Tài liệu tham khảo

[1] A. Balali, M. Asadpour, and H. Faili, "A Supervised Method to Predict the Popularity of News Articles," *Computación y Sistemas*, vol. 21, no. 4, pp. 715-726, 2017. [Online]. Available: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2737

[2] R. Bandari, S. Asur, and B. A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity," *in Proc. 6th Int. AAAI Conf. Weblogs and Social Media (ICWSM)*, Dublin, Ireland, 2012, pp. 26-33.

[3] P. Rathord, A. Jain, and C. Agrawal, "A Comprehensive Review on Online News Popularity Prediction using Machine Learning Approach," *SMART MOVES JOURNAL IJOSCIENCE*, vol. 5, no. 7, pp. 1-7, 2019.

[4] W. Stokowiec, T. Trzcinski, K. Wolk, K. Marasek, and P. Rokita, "Shallow reading with Deep Learning: Predicting popularity of online content using only its title," arXiv preprint arXiv:1707.06806, 2017. [Online]. Available: https://arxiv.org/abs/1707.06806

[5] C. Chen, P. Huang, Y. Huang and C. Lin, "Approach to Predicting News - A Precise Multi-LSTM Network With BERT," unpublished..

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[7] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. New York, NY, USA: McGraw-Hill/Irwin, 2005.

[8] T. Elmas, S. Selim, and C. Houssiaux, "Measuring and Detecting Virality on Social Media: The Case of Twitter's Viral Tweets Topic," arXiv preprint arXiv:2303.06120, 2023. [Online]. Available: https://arxiv.org/abs/2303.06120

[9] M. Glenski and T. Weninger, "Predicting User-Interactions on Reddit," arXiv preprint arXiv:1707.00195, 2017. [Online]. Available: https://arxiv.org/abs/1707.00195.