

ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo về dữ liệu

Đề tài: Hệ thống gắn nhãn bài báo tự động

Môn học: Nhập môn học máy

Sinh viên thực hiện:

Tô Hữu Danh (23127336)

Nguyễn Đăng Hưng (23127050)

Lê Phú Cường (23127164)

Nguyễn Bá Đăng Khoa (23127392)

Giáo viên hướng dẫn:

TS. Bùi Tiến Lên

Ngày 23 tháng 11 năm 2025



Mục lục

1	Giới thiệu	1
2	Phương pháp và nguồn thu thập dữ liệu	2
2.1	Nguồn dữ liệu	2
2.2	Phương pháp thu thập dữ liệu	2
2.2.1	Tìm kiếm khả năng mở rộng nguồn dữ liệu	2
2.2.2	Trích xuất dữ liệu	3
2.2.3	Tổ chức lưu trữ dữ liệu	4
2.3	Quy trình làm sạch và tiền xử lý dữ liệu	5
2.4	Đảm bảo chất lượng dữ liệu	6
2.5	Lưu trữ và quản lý dữ liệu	10
3	Phân tích khám phá dữ liệu	12
3.1	Thống kê mô tả	12
3.1.1	Quy mô và cấu trúc dữ liệu	12
3.1.2	Tỉ lệ dữ liệu thiếu	12
3.1.3	Thống kê cơ bản cho các biến số mô tả	12
3.1.4	Phân bố dữ liệu theo từng thuộc tính	12
3.2	Phân tích phân bố nhãn	13
3.2.1	Tỉ lệ nhãn theo chủ đề	13
3.2.2	Đánh giá mất cân bằng lớp	14
3.3	Phân tích mối quan hệ giữa các thuộc tính	14
3.4	Phân tích mối quan hệ giữa các thuộc tính	15
3.5	Kiểm tra chất lượng dữ liệu	19
3.6	EDA cho dữ liệu phi cấu trúc	20
4	Tài liệu tham khảo	23

Danh sách bảng

1	Bảng so sánh tỉ lệ phân bố của 3 tập dữ liệu	10
2	Bảng thống kê cơ bản	12
3	IQR cho các biến số	13
4	Bảng tỉ lệ phân bố nhãn	14

Danh sách hình vẽ

1	Kết quả trả về từ API ẩn của báo Thanh Niên	3
2	Đầu ra của bước kiểm tra giá trị rỗng. Cho thấy có 36 giá trị rỗng ở 2 cột publication_date và title	7
3	Đầu ra của bước kiểm tra trùng lặp. Cho thấy không có giá trị trùng lặp giữa các mẫu.	8
4	Biểu đồ thể hiện độ dài của các bài báo. Các bài báo trong tập dữ liệu có độ dài chủ yếu phân bố từ 3000-4000 từ.	8
5	Biểu đồ thể hiện phân bố nhãn của tập dữ liệu. Cho thấy sự phân bố đều một cách lý tưởng.	9
6	Biểu đồ thể hiện việc các bài báo có chứa mã thực thi hay không. Cho thấy không hề có mã thực thi trong dữ liệu.	9
7	Đầu ra của bước kiểm tra thủ công sau cùng. Chọn ngẫu nhiên một bài báo rồi hiện ra khoảng 300 từ đầu tiên trong bài báo. Kết quả cho thấy dữ liệu sau khi phân tích khám phá đã sạch sẽ.	10
8	Phân bố độ dài nội dung bài viết	13
9	Biểu đồ phân phối nhãn	14
10	Phân bố độ dài nội dung bài viết	15
11	Phân bố độ dài tiêu đề	16
12	Số lượng bài viết theo ngày	17
13	Số lượng bài viết theo giờ trong ngày	18
14	Heatmap số lượng bài theo Source và Main Tag	19
15	Độ dài trung bình nội dung theo Main Tag	20
16	Quan hệ giữa số câu và độ dài nội dung	21
17	Top 30 từ xuất hiện nhiều nhất sau khi lọc stopwords	21
18	Độ dài nội dung theo Main Tag	22

1 Giới thiệu

Trong kỷ nguyên số hoá hiện nay, khối lượng thông tin được tạo ra hàng ngày trên các phương tiện truyền thông trực tiếp đang tăng trưởng theo cấp số nhân. Đối với các toà soạn báo điện tử và các hệ thống tổng hợp tin tức tại Việt Nam, việc quản lý, tổ chức và phân phối nội dung đến đúng đối tượng độc giả là một thách thức lớn. Các phương pháp phân loại thủ công truyền thống không còn đáp ứng được yêu cầu về tốc độ và khả năng mở rộng, đồng thời dễ xảy ra sai sót do yếu tố chủ quan của con người. Hơn nữa, cách mà độc giả tiếp cận thông tin hiện đại đã chuyển từ việc đọc thụ động theo chuyên mục sang tìm kiếm chủ động theo từ khoá và chủ đề.

Xuất phát từ nhu cầu đó, đề án tập trung nghiên cứu và xây dựng một hệ thống phân loại văn bản tự động cho tin tức tiếng Việt. Mục tiêu cốt lõi là phát triển một giải pháp có khả năng gán nhãn chủ đề chính xác cho bài báo dựa trên nội dung văn bản, giúp tối ưu hoá trải nghiệm người dùng và hỗ trợ quá trình biên tập nội dung.

Để giải quyết bài toán phân loại văn bản tiếng Việt đa lớp (*Multi-class Text Classification*), đề án thực hiện khảo sát và triển khai so sánh hiệu năng trên ba hướng tiếp cận đại diện cho các giai đoạn phát triển của xử lý ngôn ngữ tự nhiên (*Natural Language Processing*):

- **Mô hình Học máy truyền thống (Baseline):** sử dụng phương pháp biểu diễn văn bản **TF-IDF** (*Term Frequency-Inverse Document Frequency*) kết hợp với thuật toán **Support Vector Machine**. Đây là phương pháp nền tảng, giúp thiết lập mức chuẩn về độ chính xác thời gian huấn luyện.
- **Mô hình Học sâu (Deep Learning):** Sử dụng **FastText**, một thư viện được phát triển bởi Facebook AI Research, FastText có ưu điểm vượt trội về tốc độ huấn luyện và khả năng xử lý tốt các từ hiếm (*out-of-vocabulary*) nhờ vào việc sử dụng thông tin n-gram mức thứ tự, phù hợp với đặc điểm hình thái của tiếng Việt.
- **Mô hình Ngôn ngữ Tiên huấn luyện (State-of-the-art):** Ứng dụng **PhoBERT**, một mô hình dựa trên kiến trúc Transformer (tương tự RoBERTa) được huấn luyện chuyên biệt trên dữ liệu tiếng Việt quy mô lớn. PhoBERT tận dụng cơ chế Attention để nắm bắt ngữ cảnh sâu sắc của từ ngữ, hứa hẹn mang lại độ chính xác cao nhất.

2 Phương pháp và nguồn thu thập dữ liệu

2.1 Nguồn dữ liệu

Tập dữ liệu của nhóm được thu thập thông qua việc cào hơn 20,000 bài báo từ trang báo Thanh Niên (<https://thanhnien.vn>). Thông tin mỗi bài báo bao gồm đường link, tiêu đề, ngày đăng, nội dung, nhãn chính và các nhãn phụ nếu có.

Các bài báo được cào phân bổ đều theo chủ đề chính được gán cho bài báo đó, gồm 7 tag chính: Thể thao, Thời sự, Chính trị, Kinh tế, Giáo dục, Sức khỏe, Thế giới.

2.2 Phương pháp thu thập dữ liệu

Để thu thập được số lượng lớn bài báo mà vẫn đảm bảo phân loại chính xác chủ đề, nhóm cần giải quyết được 3 vấn đề chính:

- Khả năng mở rộng nguồn dữ liệu, tức là tự động hoá việc lấy thêm đường link các bài báo.
- Phân tích cấu trúc HTML của một bài báo và trích xuất thông tin như tag, tiêu đề, nội dung,...
- Tổ chức lưu trữ dữ liệu đã thu thập được sao cho dễ truy xuất và mở rộng.

2.2.1 Tìm kiếm khả năng mở rộng nguồn dữ liệu

Hầu hết các trang báo mạng sẽ có các chuyên mục để người dùng có thể đọc thể loại mình muốn. Bằng việc giới hạn phạm vi thu thập trong các chuyên mục định sẵn, nhóm sẽ giải quyết được bài toán gán nhãn dữ liệu và giảm bớt công việc trong quá trình làm sạch.

Một vấn đề với cách tiếp cận này các cơ chế phân trang truyền thống trên VnExpress hay Dân Trí thường áp đặt giới hạn cứng ở mức 20-30 trang lịch sử, gây khó khăn cho việc xây dựng tập dữ liệu lớn.

Tuy nhiên, thông qua kỹ thuật phân tích gói tin HTTP bằng công cụ **Burp Suite**, nhóm đã xác định được cơ chế truy xuất dữ liệu thông qua API ẩn của báo Thanh Niên. Khi nhấn vào nút "Xem Thêm", thay vì render lại toàn bộ trang web với nội dung mới, client sử dụng endpoint `GET timelinelist/{id}/{page_number}.htm` để trả về dữ liệu thô chứa đường link và hình ảnh của các bài báo trong một trang lịch sử và render nó ở cuối. Điểm đặc biệt nằm ở việc tham số

page_number có thể được tùy biến theo ý muốn, vượt qua con số 20 - 30, cho phép nhóm tự động hóa việc truy xuất các bài viết cũ nằm ngoài phạm vi hiển thị mặc định của giao diện người dùng.



Hình 1: Kết quả trả về từ API ẩn của báo Thanh Niên

2.2.2 Trích xuất dữ liệu

Nhóm sử dụng 3 thư viện Python cho việc trích xuất dữ liệu:

- **requests**: Thư viện dùng để gửi HTTP requests tới API cũng như đến các bài báo
- **Beautiful Soup**: Thư viện dùng để trích xuất thông tin từ dạng HTML, XML,.. theo điều kiện mong muốn.
- **sqlite3**: Connector tới sqlite, một hệ quản trị cơ sở dữ liệu SQL nhẹ, serverless, các database sẽ được xuất ra thành từng file, dùng cho việc lưu trữ các bài báo

Quy trình trích xuất dữ liệu từ một bài báo của nhóm gồm:

- Lặp qua các id của từng chuyên mục và số lượng page theo cài đặt (150 trong lần chạy của nhóm).
- Với mỗi id và `page_number`, gửi một HTTP request tới API ẩn của báo Thanh Niên, dùng `Beautiful Soup` lọc ra danh sách các đường link đến bài báo
- Gửi request tới đường link của từng bài báo đó nhận nội dung HTML trả về.
- Tiếp tục dùng `Beautiful Soup` để duyệt qua các HTML tag chứa nội dung bài báo gồm tiêu đề, ngày đăng, nội dung, các tag phụ và trích xuất chúng.
- Thực hiện lưu trữ nội dung đã trích xuất vào database.

2.2.3 Tổ chức lưu trữ dữ liệu

Nhóm đã sử dụng SQLite để tổ chức lưu trữ dữ liệu. Toàn bộ dữ liệu sau khi cào được nằm trong một file `articles.sqlite3` với 2 bảng `articles` và `sub_tags`

```
1 CREATE TABLE articles (  
2     link TEXT PRIMARY KEY,  
3     publication_date DATETIME,  
4     content TEXT,  
5     main_tag TEXT,  
6     source TEXT  
7 );  
8  
9 CREATE TABLE sub_tags (  
10    link TEXT,  
11    tag TEXT,  
12    PRIMARY KEY (link, tag),  
13    FOREIGN KEY(link) REFERENCES articles(link)  
14 );
```

Listing 1: Mã nguồn tạo bảng CSDL

Nhóm quyết định dùng mô hình cơ sở dữ liệu quan hệ (relational database) thay vì dùng file JSON hay CSV vì các lý do sau đây:

- CSV không thích hợp để chứa nội dung các bài báo, vốn gồm nhiều dấu phẩy (,) và dấu nháy kép (").

- Việc cập nhật dữ liệu vào file JSON truyền thống gặp khó khăn lớn khi kích thước file tăng lên, do quy trình này thường đòi hỏi phải xử lý lại toàn bộ nội dung file để đảm bảo đúng cấu trúc cú pháp.
- Với SQLite, nhóm có thể thiết lập quy trình ghi dữ liệu liên tục theo từng lô nhỏ (5 bài một lần insert) để đảm bảo hiệu suất mà có thể tránh các rủi ro kỹ thuật (mất mạng, sập nguồn) làm mất data trong quá trình cào bài báo.
- Việc dùng CSDL đã có ép kiểu (DATETIME) và đặt khoá chính từ trước sẽ giúp loại bỏ các dữ liệu trùng lặp hoặc gặp lỗi trong quá trình trích xuất, giúp làm giảm công việc cho khâu làm sạch.

2.3 Quy trình làm sạch và tiền xử lý dữ liệu

Trước khi đưa vào bất cứ mô hình nào, dữ liệu thô đều trải qua các bước làm sạch cơ bản để loại bỏ nhiễu:

- **Xử lý giá trị thiếu và trùng lặp:** Thực hiện loại bỏ các mẫu dữ liệu bị khuyết thiếu thông tin quan trọng (như tiêu đề hoặc nội dung) và các bài viết bị trùng lặp hoàn toàn để tránh hiện tượng rò rỉ dữ liệu (*data leakage*) giữa các tập huấn luyện và kiểm thử. Sau bước này, kích thước bộ dữ liệu còn lại là 20,777 bài viết.
- **Loại nhiễu định dạng:** Dữ liệu văn bản được quét để phát hiện và loại bỏ các đoạn mã HTML, CSS hoặc JavaScript không mang ngữ nghĩa, đảm bảo nội dung đầu vào hoàn toàn là văn bản tự nhiên.
- **Chuẩn hoá bảng mã Unicode:** Do đặc thù tiếng Việt có hai kiểu gõ phổ biến (dạng sẵn và tổ hợp), toàn bộ văn bản được chuẩn hoá về định dạng Unicode NFC để đảm bảo tính nhất quán trong biểu diễn ký tự.

Tuy nhiên, các mô hình có sự khác biệt đặc thù về cơ chế hoạt động, chủ yếu là sự khác biệt của Transformer so với phần còn lại. Vì thế giai đoạn tiền xử lý được chia thành hai luồng riêng biệt:

- Đối với các mô hình cơ bản (TF-IDF + SVM, FastText), việc giảm chiều dữ liệu và thống nhất từ vựng là ưu tiên hàng đầu:

- **Chuẩn hoá ký tự thường (*Lowercasing*):** Chuyển toàn bộ văn bản về chữ thường để giảm kích thước bộ từ điển mà không làm mất đi quá nhiều ngữ nghĩa quan trọng.
 - **Tách từ (*Word Segmentation*):** Sử dụng thư viện `pyvi` để gộp các từ ghép tiếng Việt, giúp nhận diện đúng đơn vị từ vựng.
 - **Loại bỏ từ dừng (*Stopword Removal*):** Loại bỏ các hư từ (*stopwords*) dựa trên danh sách từ dừng tiếng Việt. Bước này giúp giảm nhiễu và tập trung vào các từ mang thông tin chính trị, kinh tế, xã hội đặc thù.
- Đối với mô hình ngôn ngữ tiền huấn luyện PhoBERT, ngữ cảnh và cấu trúc câu đóng vai trò then chốt. Do đó, chiến lược xử lý được điều chỉnh như sau:
 - **Giữ nguyên định dạng (*No Lowercasing/Stopword Removal*):** Chúng tôi không chuyển văn bản về chữ thường và không loại bỏ từ dừng. Việc này nhằm bảo toàn cấu trúc ngữ pháp và thông tin về tên riêng, giúp cơ chế *Attention* của PhoBERT hoạt động hiệu quả nhất.
 - **Tách từ (*Word Segmentation*):** Tương tự như luồng cơ bản, văn bản đầu vào cho PhoBERT cũng cần được tách từ để đồng bộ với dữ liệu mà PhoBERT đã được huấn luyện trước đó.

Cuối cùng, các nhãn đầu ra phải được mã hoá bằng cách ánh xạ các nhãn (gồm 7 nhãn cho ngữ cảnh của đề án này) thành các con số nguyên (từ 0 tới 6 cho ngữ cảnh của đề án này) để phục vụ quá trình tính toán của máy.

2.4 Đảm bảo chất lượng dữ liệu

Để đảm bảo mô hình phân loại đạt hiệu suất cao và có khả năng tổng quát hoá tốt, bộ dữ liệu huấn luyện cần thoả mãn các tiêu chí:

- **Tính đầy đủ:** Các mẫu dữ liệu không được khuyết thiếu các trường thông tin quan trọng.
- **Tính sạch:** Dữ liệu văn bản không chứa các ký tự lạ, mã lỗi (HTML tags, JavaScript) hoặc các nội dung rác (quảng cáo hay link spam).
- **Tính nhất quán:** Dữ liệu phải đồng nhất về định dạng (Unicode NFC), ngôn ngữ (tiếng Việt) và cách gán nhãn chủ đề.

- **Tính cân bằng:** Số lượng mẫu dữ liệu giữa các nhãn không được chênh lệch quá lớn để tránh hiện tượng mô hình bị thiên vị (*bias*) về phía lớp đa số.

Quy trình kiểm tra gồm 2 bước chính:

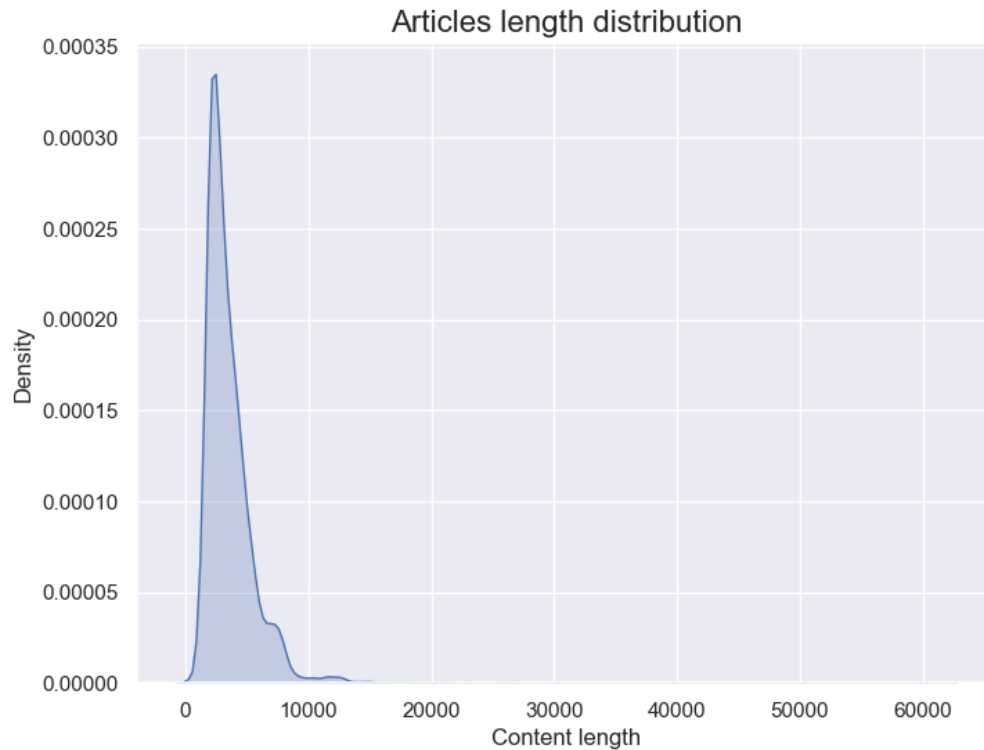
- **Kiểm tra tự động (*Automatic checks*):** Dựa trên các tiêu chí đã đề ra, xây dựng script để kiểm tra tự động quy trình EDA:
 - **Kiểm tra giá trị rỗng:** Sử dụng pandas để quét toàn bộ DataFrame. Kết quả phát hiện 36 mẫu bị thiếu thông tin nghi là các quảng cáo, các mẫu đã được loại bỏ tự động.
 - **Kiểm tra trùng lặp:** Rà soát dựa trên đường dẫn và nội dung bài báo để đảm bảo tính duy nhất.
 - **Kiểm tra độ dài văn bản:** Tính toán độ dài ký tự của từng bài báo để phát hiện các bài quá ngắn hoặc quá dài.
 - **Kiểm tra phân bố nhãn:** Thống kê số lượng bài viết theo từng chủ đề. Kết quả cho thấy dữ liệu đạt độ cân bằng rất tốt, không cần áp dụng các kỹ thuật tái lấy mẫu (*Resampling*).
- **Kiểm tra thủ công (*Manual checks*):** Lấy ngẫu nhiên 5 bài viết từ các nhãn khác nhau. Đọc lướt nội dung để xác minh nhãn chủ đề ban đầu có khớp với nội dung thực tế không. Kết quả là dữ liệu đạt độ sạch hoàn hảo.

```
link      0
publication_date  36
title     36
content   0
main_tag  0
source    0
dtype: int64
```

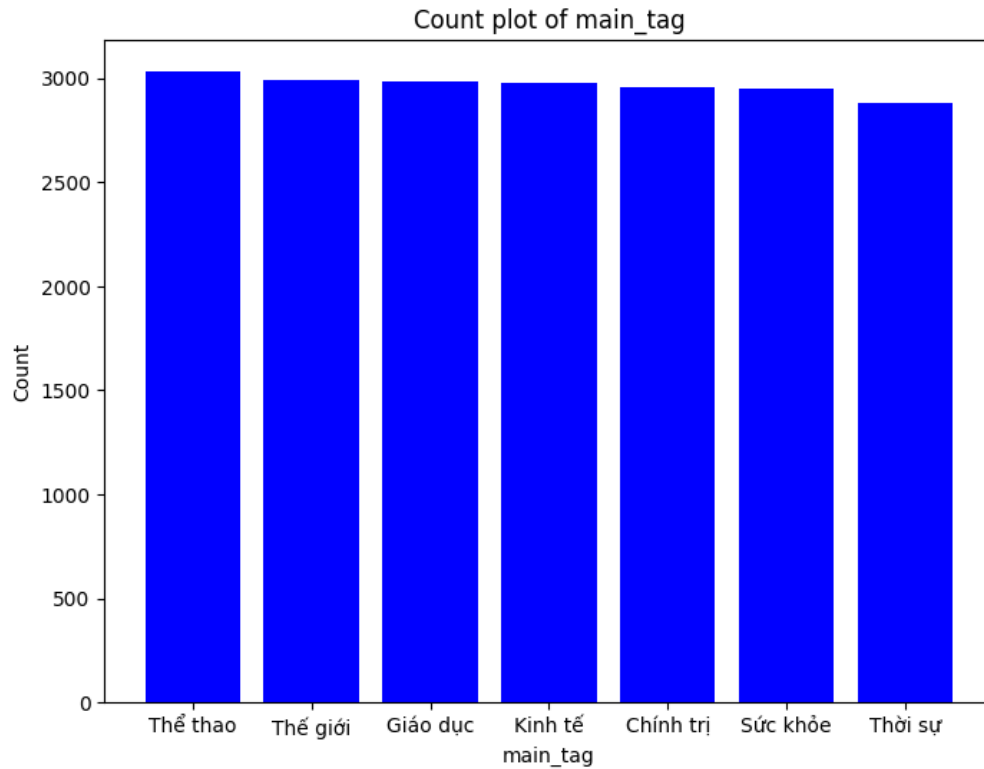
Hình 2: Đầu ra của bước kiểm tra giá trị rỗng. Cho thấy có 36 giá trị rỗng ở 2 cột `publication_date` và `title`.

```
link      20777
publication_date  19456
title     20755
content   20777
main_tag    7
source     1
dtype: int64
```

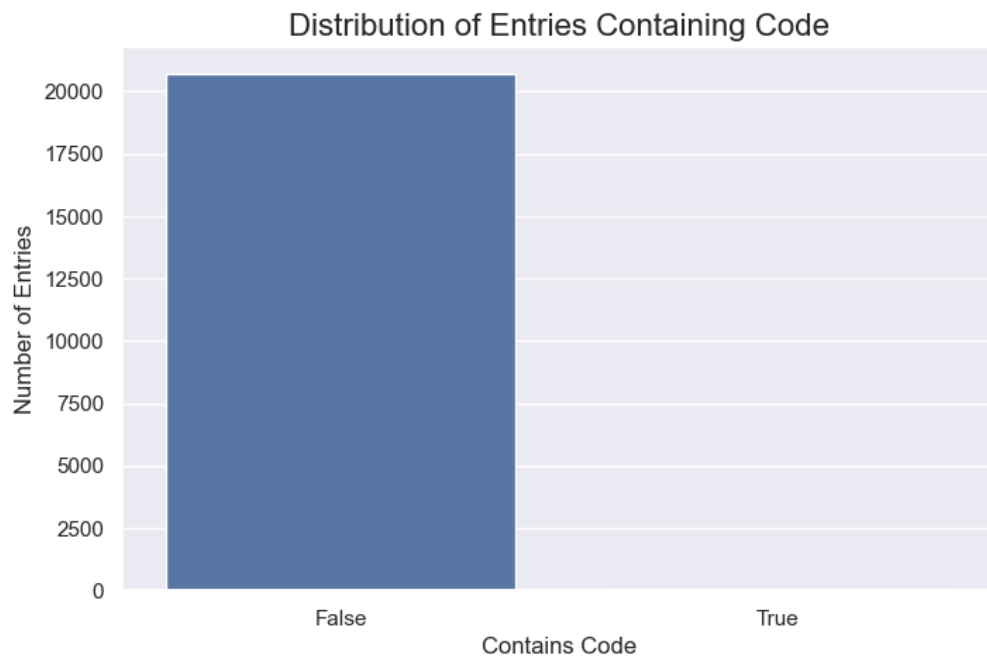
Hình 3: Đầu ra của bước kiểm tra trùng lặp. Cho thấy không có giá trị trùng lặp giữa các mẫu.



Hình 4: Biểu đồ thể hiện độ dài của các bài báo. Các bài báo trong tập dữ liệu có độ dài chủ yếu phân bố từ 3000-4000 từ.



Hình 5: Biểu đồ thể hiện phân bố nhãn của tập dữ liệu. Cho thấy sự phân bố đều một cách lý tưởng.



Hình 6: Biểu đồ thể hiện việc các bài báo có chứa mã thực thi hay không. Cho thấy không hề có mã thực thi trong dữ liệu.

```
=====
DEMO KẾT QUẢ TIỀN XỬ LÝ DỮ LIỆU (MẪU NGẪU NHIÊN)
=====
1. THÔNG TIN CHUNG:
- Tiêu đề: Kiến nghị tăng doanh thu hộ kinh doanh lên 2-3 tỉ đồng mới xuất hóa đơn
- Ngày đăng: 01/08/2025 13:33 GMT+7
- Chuyên mục: Kinh tế (Label ID: N/A)
- Link gốc: https://thanhnien.vn/kiem-ghi-tang-doanh-thu-ho-kinh-doanh-len-2-3-ti-dong-moi-xuat-hoa-don-185250801132809942.htm

2. SO SÁNH QUY TRÌNH XỬ LÝ (Hiển thị 300 ký tự đầu):

[A] Dữ liệu thô (Raw/Standardized):
Ngày 1.8, tại hội thảo lấy ý kiến về định hướng chính sách của dự thảo Luật Quản
lý thuế (thay thế) do Cục Thuế (Bộ Tài chính) tổ chức, đại diện Liên đoàn Thương
mại và Công nghiệp Việt Nam (VCCI) kiến nghị tăng ngưỡng áp dụng hóa đơn điện tử
từ mức tính tiền kết nối cơ quan thuế (HĐĐT KNCQT) từ 1 t...
```

Hình 7: Đầu ra của bước kiểm tra thủ công sau cùng. Chọn ngẫu nhiên một bài báo rồi hiện ra khoảng 300 từ đầu tiên trong bài báo. Kết quả cho thấy dữ liệu sau khi phân tích khám phá đã sạch sẽ.

2.5 Lưu trữ và quản lý dữ liệu

Để đảm bảo khả năng tái lập (*reproducibility*) và thuận tiện cho việc huấn luyện trên nhiều kiến trúc mô hình khác nhau, dữ liệu được quản lý theo quy trình 3 tầng:

- **Dữ liệu trung gian:** Là dữ liệu sau bước phân tích khám phá, được lưu dưới dạng Pickle (.pkl) để giữ nguyên các thuộc tính đối tượng Python (như DataFrame) sau khi làm sạch sơ bộ, giúp tiết kiệm thời gian tải dữ liệu của quá trình kế tiếp.
- **Dữ liệu huấn luyện:** Được chia tách và lưu trữ dưới dạng CSV trong các thư mục riêng biệt, sẵn sàng nạp vào mô hình.

Tập dữ liệu sau cùng sẽ được phân tách thành 3 tập riêng biệt là Train, Validation và Test với tỉ lệ 80%-10%-10%. Tuy nhiên cần áp dụng kỹ thuật lấy mẫu phân tầng (*Stratified Sampling*) để đảm bảo phân bố đều các nhãn ra các tập dữ liệu. Dưới đây là bảng phân bố các nhãn:

label_encoded	Train	Validation	Test
0	0.145960	0.145813	0.145813
1	0.143854	0.143888	0.143888
2	0.143794	0.143407	0.143888
3	0.143132	0.143407	0.142926
4	0.142410	0.142445	0.142445
5	0.142109	0.141963	0.142445
6	0.138740	0.139076	0.138595

Bảng 1: Bảng so sánh tỉ lệ phân bố của 3 tập dữ liệu

Dễ thấy từ bảng trên, tỉ lệ phân bố các nhãn của 3 tập dữ liệu khá đồng đều chứng tỏ quá trình phân chia dữ liệu đảm bảo được tính cân bằng.

3 Phân tích khám phá dữ liệu

3.1 Thống kê mô tả

3.1.1 Quy mô và cấu trúc dữ liệu

Tập dữ liệu ban đầu sau khi thu thập gồm **20,813** bài báo với 6 thuộc tính gốc: `link`, `publication_date`, `title`, `content`, `main_tag`, `source`. Sau bước loại bỏ dữ liệu thiếu quan trọng, tập dữ liệu còn lại là **20,777** bài báo. Trong quá trình EDA, nhóm bổ sung thêm một thuộc tính số là `content_length` (độ dài văn bản) để phục vụ thống kê, do đó DataFrame cuối cùng có tổng cộng 7 cột.

3.1.2 Tỷ lệ dữ liệu thiếu

Kiểm tra giá trị rỗng được thực hiện trên toàn bộ DataFrame. Kết quả cho thấy có **36** mẫu bị thiếu dữ liệu ở hai trường `publication_date` và `title` ở tập dữ liệu thô. Các mẫu này nghi là quảng cáo hoặc bài lỗi crawl nên đã được loại bỏ tự động. Sau làm sạch, tập dữ liệu dùng cho huấn luyện không còn giá trị rỗng.

3.1.3 Thống kê cơ bản cho các biến số mô tả

Do dữ liệu là văn bản phi cấu trúc, nhóm xây dựng các biến số mô tả độ dài gồm: `content_length` (độ dài nội dung theo ký tự), `content_words` (độ dài theo số từ), và các biến tương tự cho tiêu đề. Các thống kê cơ bản (mean, median, variance, min/max và IQR) được tính bằng `describe()` và `quantile()`.

Bảng 2: Bảng thống kê cơ bản

Biến	Count	Mean	Std	Min	25%	50%	75%	Max
title_length	20777	63.06	14.88	9	53	64	73	164
content_length	20777	3525.38	2106.83	104	2208	2988	4242	61863
title_words	20777	13.97	3.11	2	12	14	16	34
content_words	20777	768.37	460.55	24	480	651	927	13592

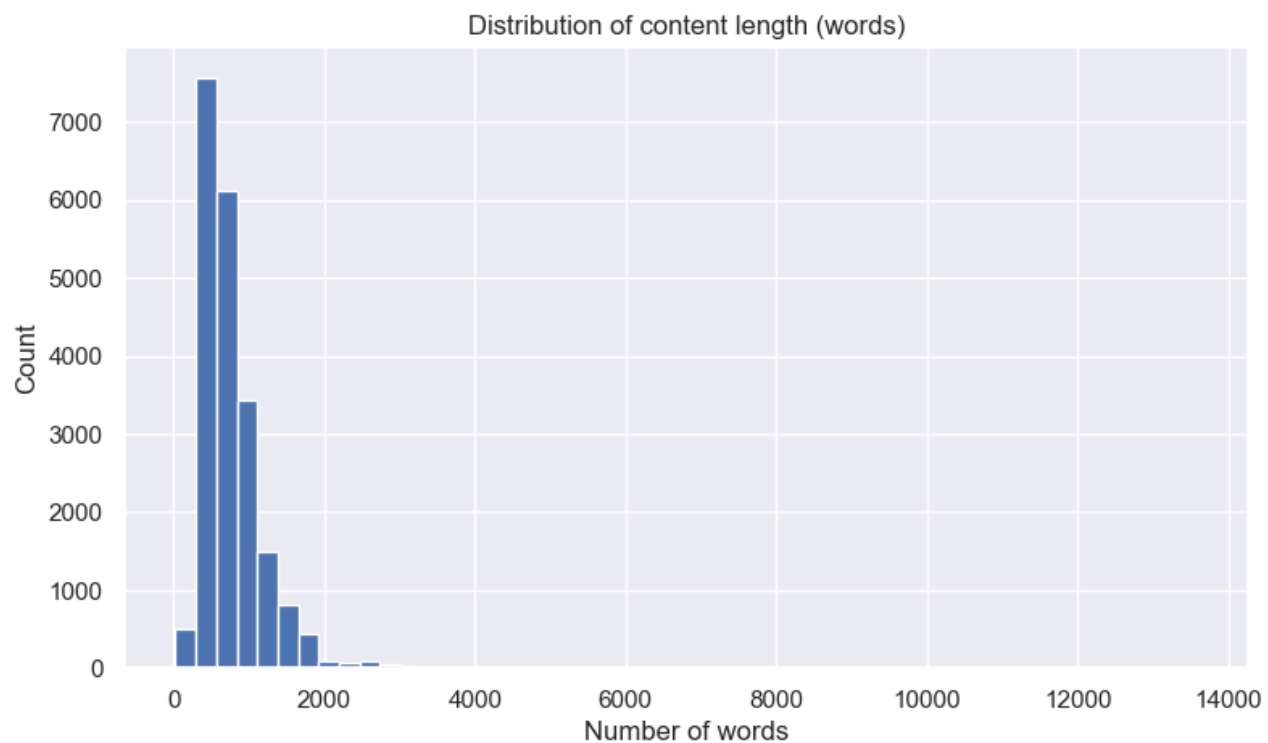
3.1.4 Phân bố dữ liệu theo từng thuộc tính

Phân bố độ dài nội dung được biểu diễn bằng đồ thị KDE/Histogram và Boxplot, cho thấy dữ liệu không có hiện tượng lệch quá mạnh hay nhiễu bất thường. Ngoài ra, dữ liệu được thu thập từ cùng

Bảng 3: IQR cho các biến số

Đặc trưng	Giá trị
title_length	20
content_length	2034
title_words	4
content_words	447

một nguồn báo (`source = Thanh Niên`), đảm bảo tính nhất quán về phong cách và ngôn ngữ.



Hình 8: Phân bố độ dài nội dung bài viết

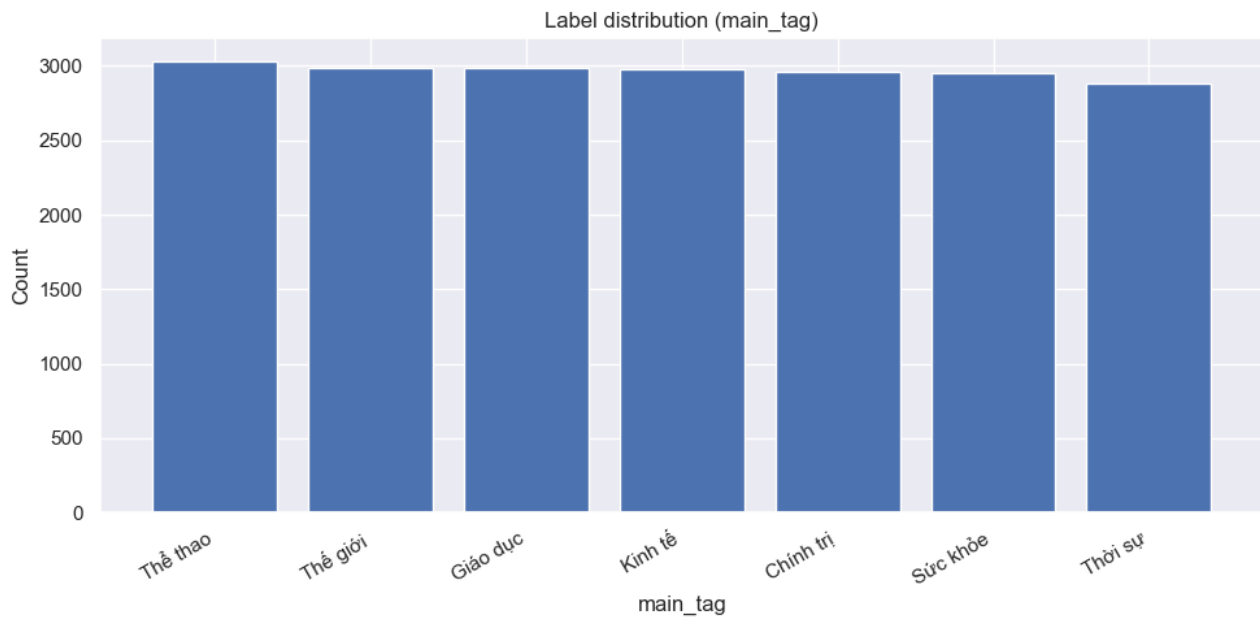
3.2 Phân tích phân bố nhãn

3.2.1 Tỷ lệ nhãn theo chủ đề

Tập dữ liệu gồm 7 nhãn chính: *Thể thao*, *Thời sự*, *Chính trị*, *Kinh tế*, *Giáo dục*, *Sức khỏe*, *Thế giới*. Nhìn vào biểu đồ và bảng phân bố nhãn cho thấy số lượng mẫu giữa các nhãn được phân bố rất đều.

Bảng 4: Bảng tỉ lệ phân bố nhãn

Chuyên mục	Giá trị
Thể thao	14.593
Thể giới	14.386
Giáo dục	14.376
Kinh tế	14.314
Chính trị	14.242
Sức khỏe	14.213
Thời sự	13.876



Hình 9: Biểu đồ phân phối nhãn

3.2.2 Đánh giá mất cân bằng lớp

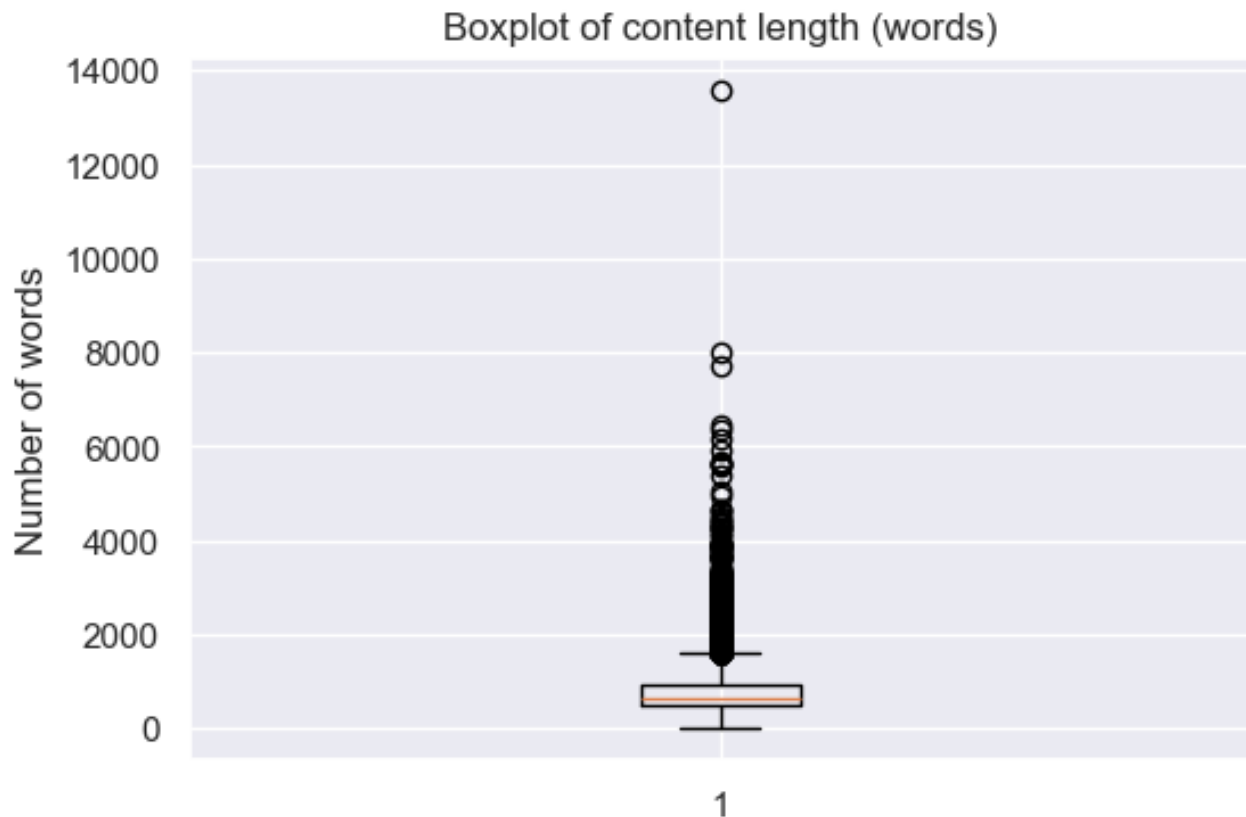
Với phân bố nhãn đồng đều như trên, tập dữ liệu **không gặp vấn đề mất cân bằng lớp (*class imbalance*)**. Do đó, nhóm không cần áp dụng các kỹ thuật tái lấy mẫu (*oversampling/undersampling*) và có thể huấn luyện mô hình trực tiếp trên phân phối hiện tại. Trong đánh giá, nhóm vẫn ưu tiên sử dụng Macro-F1 nhằm đảm bảo chất lượng trên tất cả các nhãn.

3.3 Phân tích mối quan hệ giữa các thuộc tính

Nhóm xem xét mối liên hệ giữa các thuộc tính mô tả độ dài (như `content_length`, `content_words`, `title_length`, `title_words`). Ma trận tương quan cho thấy các biến đo độ dài nội dung có tương

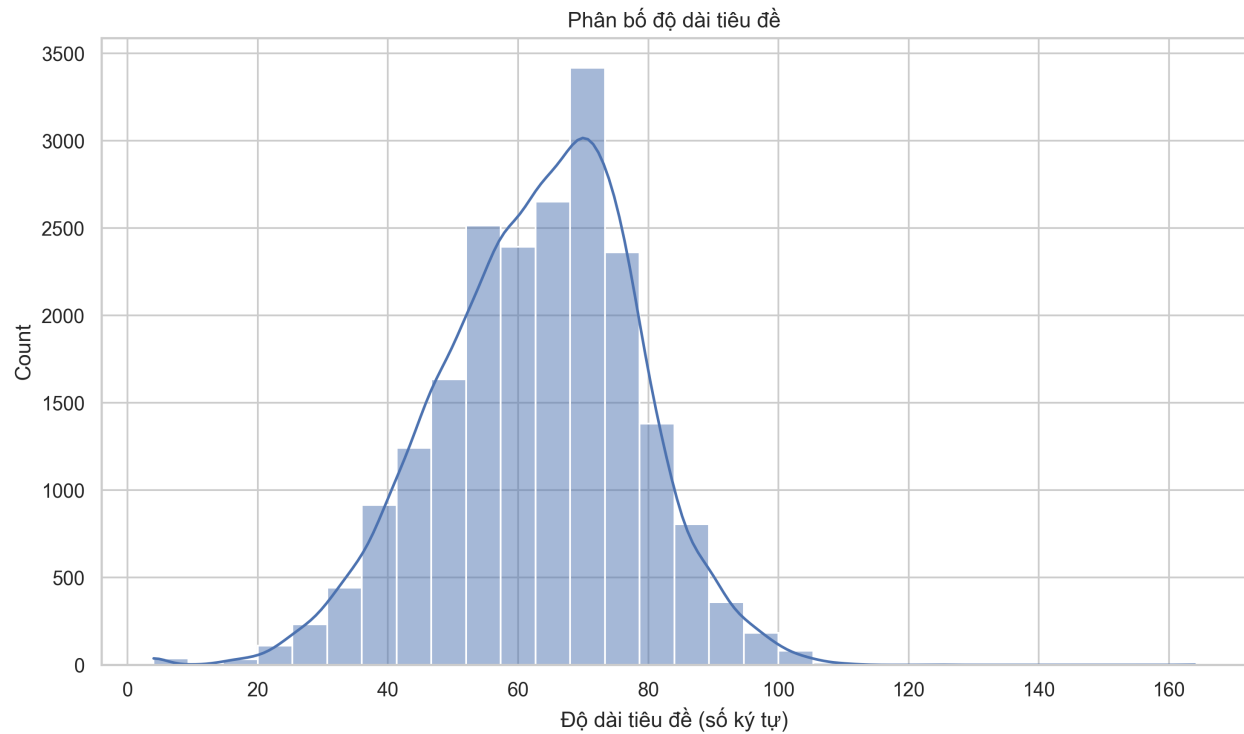
quan cao với nhau, trong khi độ dài tiêu đề chỉ tương quan yếu với độ dài nội dung. Điều này phù hợp với trực giác: bài càng dài thì số từ càng nhiều, nhưng tiêu đề không phản ánh trực tiếp độ dài bài báo.

3.4 Phân tích mối quan hệ giữa các thuộc tính



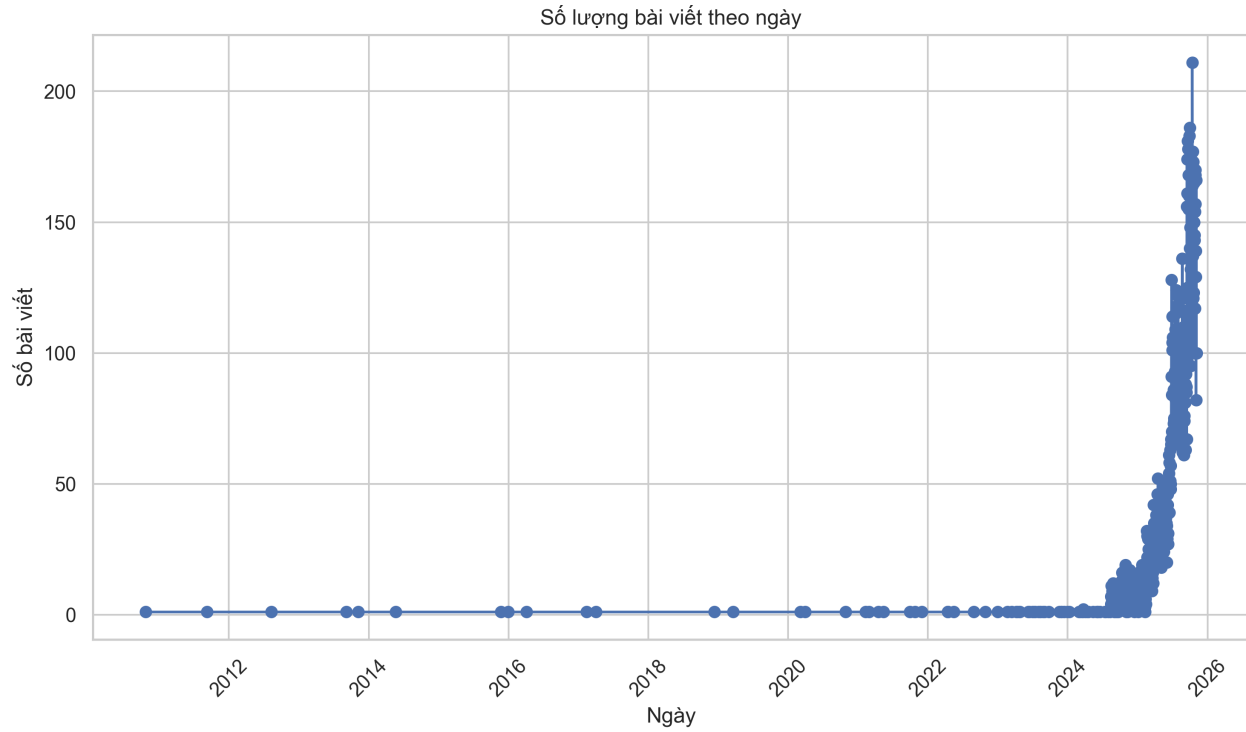
Hình 10: Phân bố độ dài nội dung bài viết

Biểu đồ Histogram cho thấy độ dài nội dung bài viết có phân bố lệch trái mạnh. Phần lớn bài báo có độ dài từ khoảng 1.000 đến 6.000 ký tự, trong khi một số ít bài có độ dài rất lớn (lên tới hơn 60.000 ký tự), tạo ra các *giá trị ngoại lệ* (outliers). Điều này phản ánh sự đa dạng về cấu trúc bài viết, từ các bản tin ngắn đến các bài phân tích chuyên sâu.



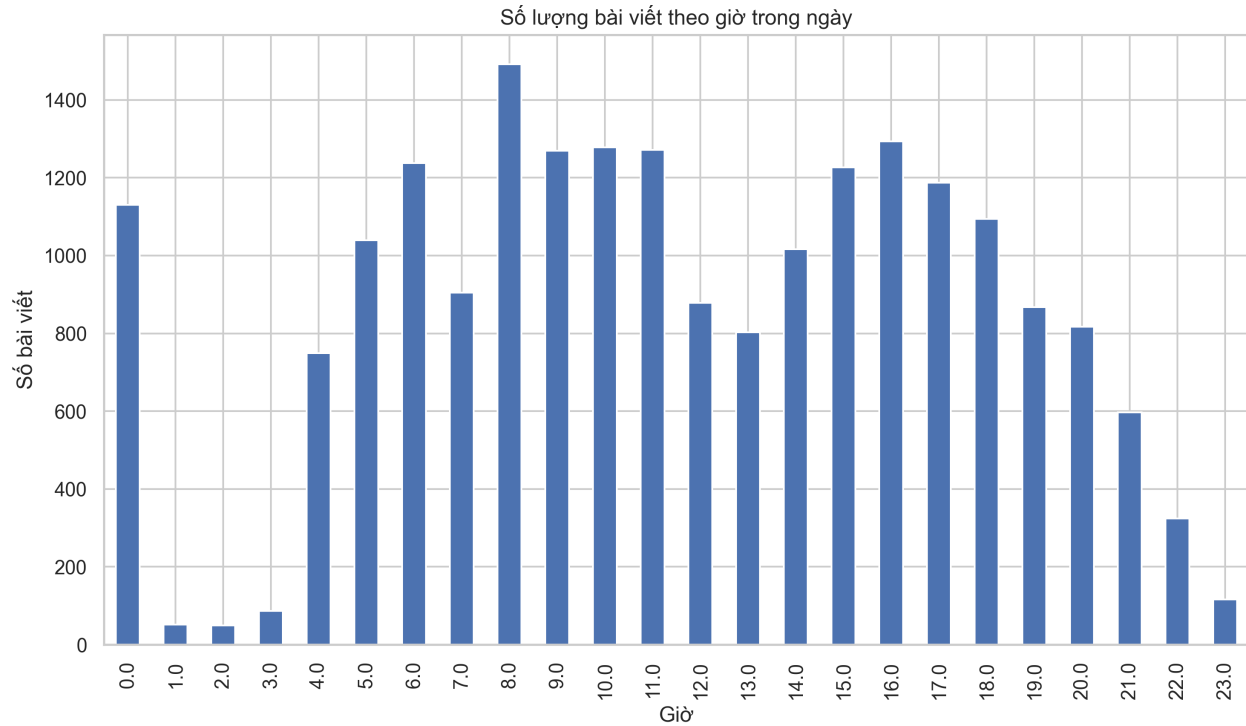
Hình 11: Phân bố độ dài tiêu đề

Độ dài tiêu đề có phân bố gần giống chuẩn, tập trung chủ yếu trong khoảng 50 đến 75 ký tự. Điều này cho thấy các bài báo thường được tối ưu hoá tiêu đề để vừa ngắn gọn, vừa truyền tải đủ thông tin với một số ít tiêu đề có độ dài cực lớn hoặc cực nhỏ xuất hiện như các trường hợp ngoại lệ.



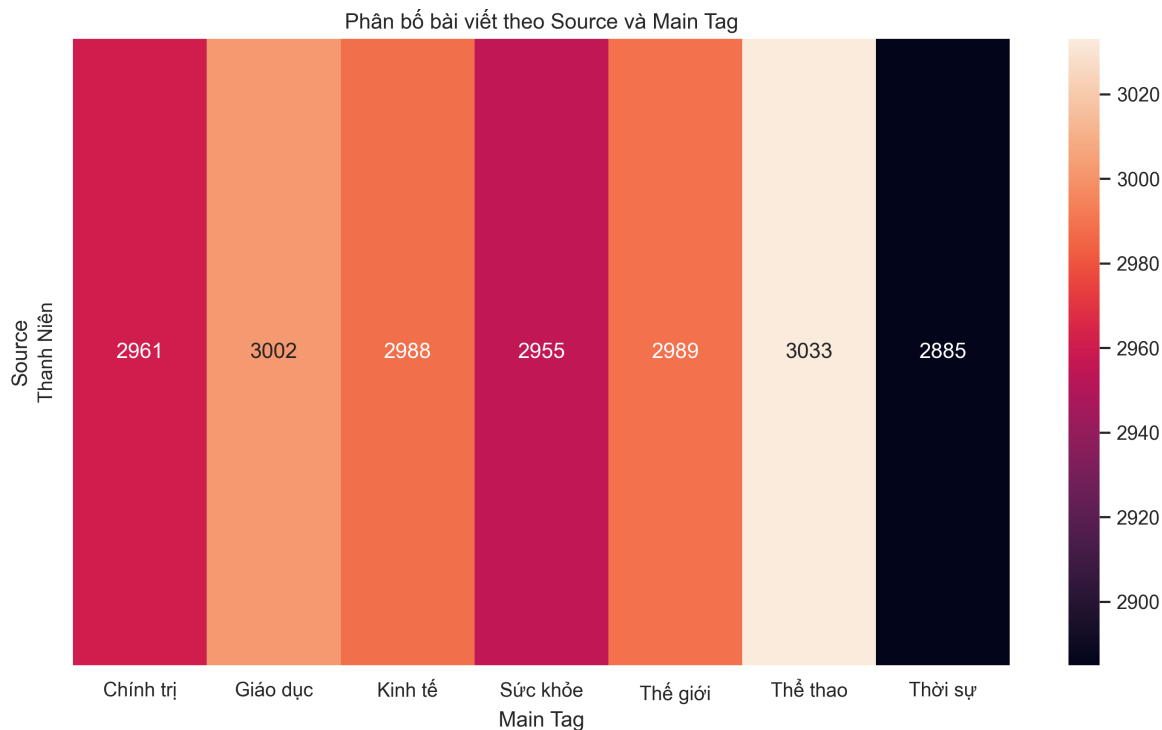
Hình 12: Số lượng bài viết theo ngày

Do dữ liệu được crawl chủ yếu trong các tháng gần đây nên số lượng bài viết tập trung trong nửa cuối năm 2025.



Hình 13: Số lượng bài viết theo giờ trong ngày

Biểu đồ cho thấy các bài báo chủ yếu được xuất bản trong khoảng 5h–23h hằng ngày. – Khung 8h–11h là thời điểm có lượng bài cao nhất (đỉnh khoảng 1.400–1.500 bài). – Tối từ 16h–20h cũng có lượng bài cao. – Giờ khuya (0h–3h) có số bài rất ít.



Hình 14: Heatmap số lượng bài theo Source và Main Tag

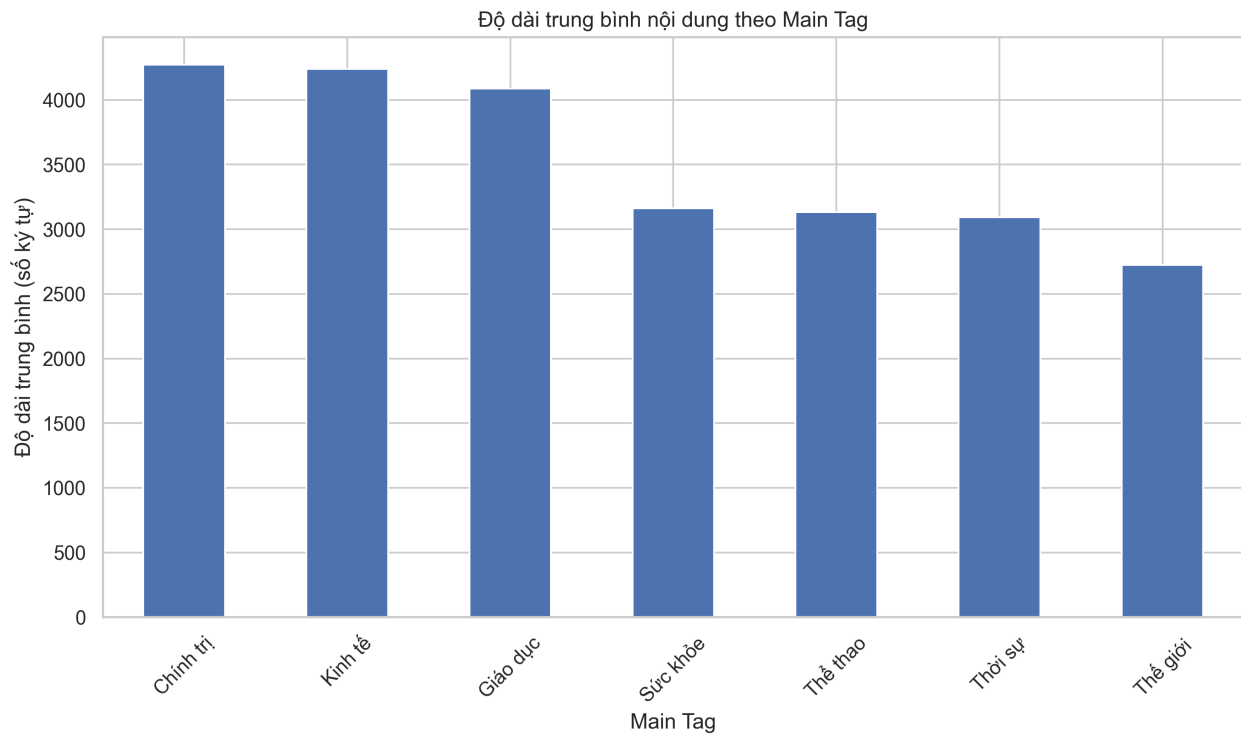
Vì toàn bộ dữ liệu đều đến từ một nguồn (“Thanh Niên”), *heatmap* cho thấy số bài giữa các `main_tag` khá đồng đều: mỗi nhóm có khoảng 2.900–3.000 bài viết. Điều này chứng tỏ bộ dữ liệu được thu thập tương đối cân bằng theo từng chủ đề.

3.5 Kiểm tra chất lượng dữ liệu

Qua phân tích phân bố: – Cả tiêu đề và nội dung đều có phân bố hợp lý, không có dấu hiệu lỗi mã hóa. – Tuy nhiên, một số bài có nội dung quá dài (60.000+ ký tự), đây có thể là bài dạng phỏng vấn hoặc tổng hợp nhiều đoạn, nhưng cũng có thể là lỗi trùng lặp hoặc gộp nội dung. – Không thấy hiện tượng tiêu đề rỗng hoặc content rỗng qua biểu đồ.

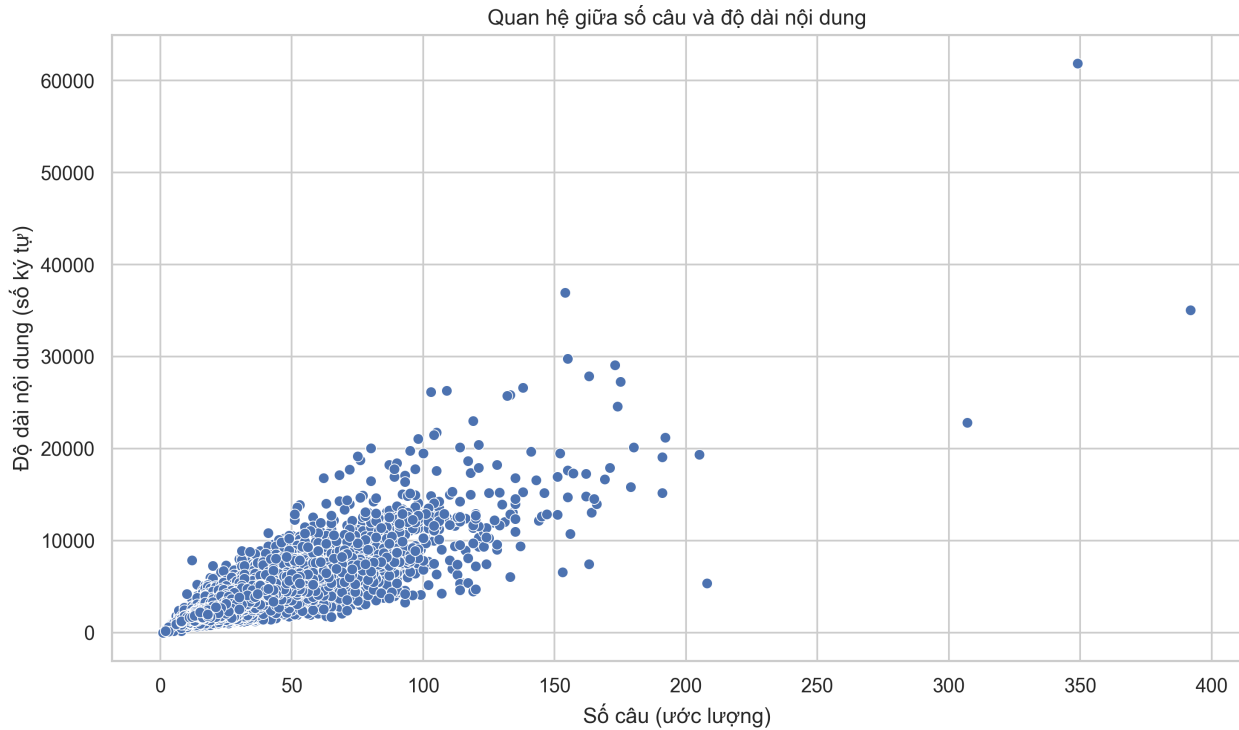
Biểu đồ phân bố theo ngày cho thấy dữ liệu lịch sử trước năm 2023 gần như không có, nhưng đây không phải lỗi mà là đặc điểm của nguồn dữ liệu (crawler chỉ hoạt động trong một số năm gần đây).

3.6 EDA cho dữ liệu phi cấu trúc



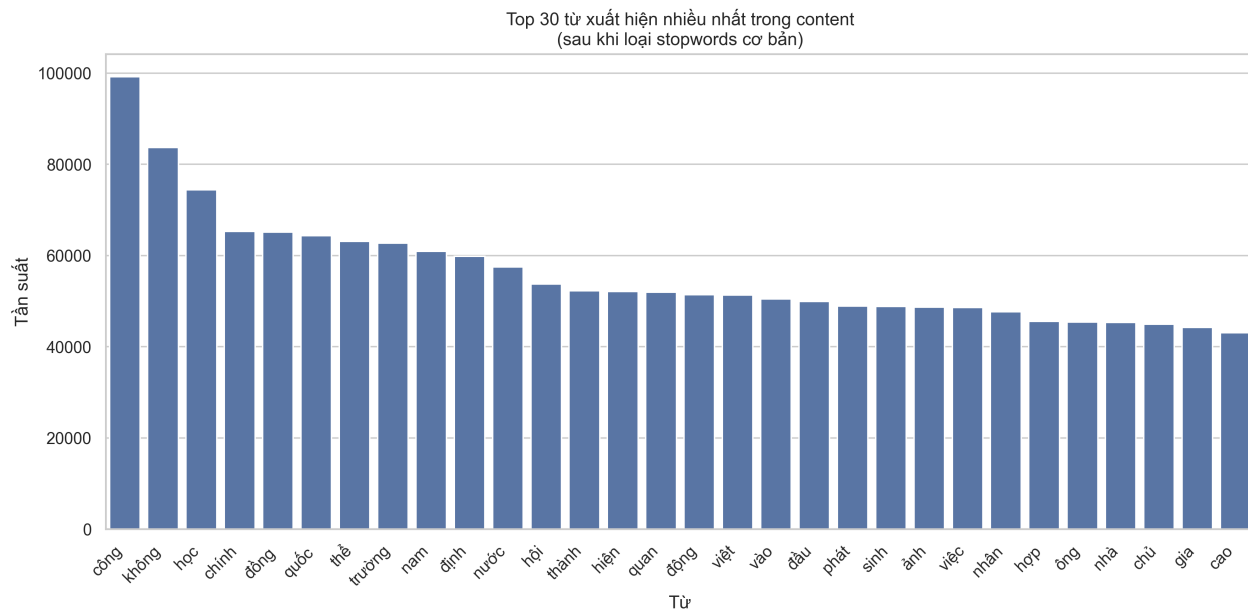
Hình 15: Độ dài trung bình nội dung theo Main Tag

Chính trị và Kinh tế có độ dài 4200 ký tự, cao nhất; Thế giới thấp nhất (2700 ký tự).



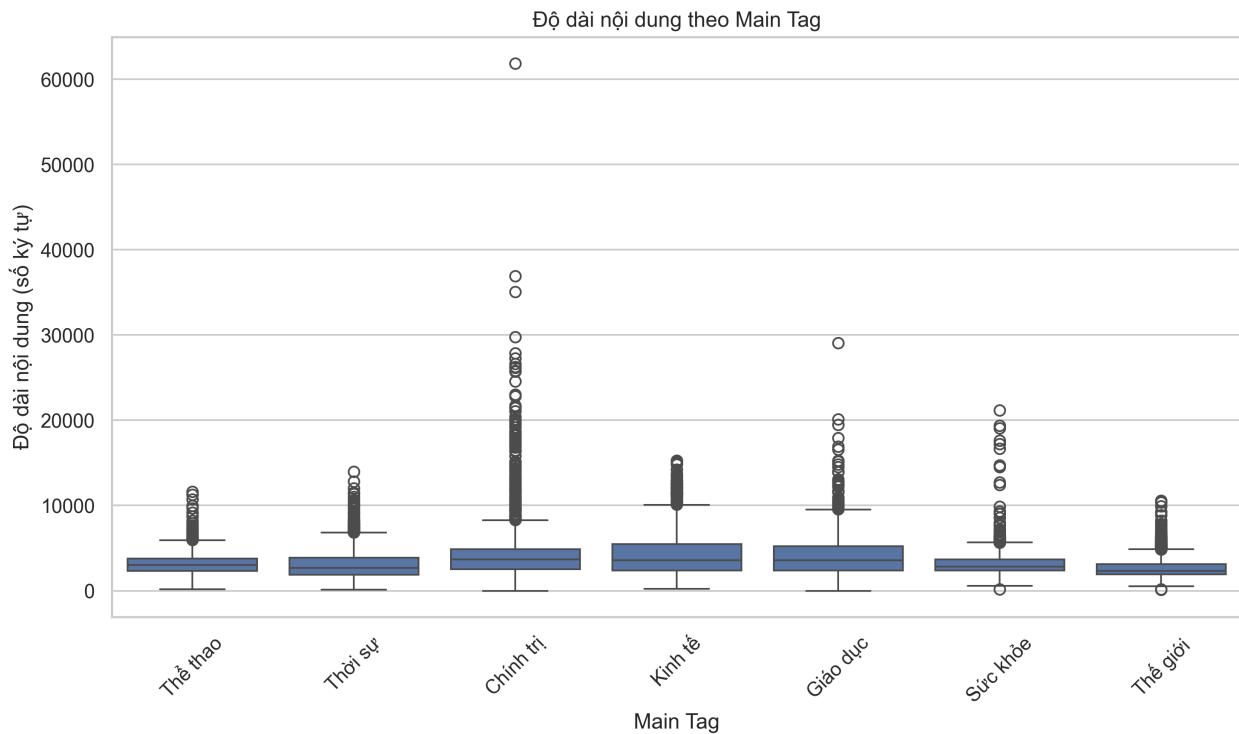
Hình 16: Quan hệ giữa số câu và độ dài nội dung

Scatterplot cho thấy mối quan hệ tuyến tính mạnh giữa số câu và độ dài nội dung: bài viết dài thường có số câu nhiều tương ứng. Phần lớn bài nằm dưới 150 câu, tuy nhiên có một số trường hợp đặc biệt lên đến 300–400 câu.



Hình 17: Top 30 từ xuất hiện nhiều nhất sau khi lọc stopwords

Các từ xuất hiện nhiều nhất gồm: “công”, “không”, “học”, “chính”, “đồng”, “quốc”, “thể”, “trường”... Các từ này chủ yếu là danh từ hoặc các từ mang nghĩa nội dung quan trọng, phản ánh tính chất báo chí của tập dữ liệu. Tần suất từ cao (40.000–100.000 lần) chứng tỏ bộ dữ liệu rất lớn và có sự lặp lại của nhiều chủ đề tương đồng.



Hình 18: Độ dài nội dung theo Main Tag

Nhóm Chính trị và Kinh tế có độ dài trung bình cao nhất (khoảng trên 4.200 ký tự), tiếp theo là Giáo dục. Các nhóm Sức khỏe, Thế giới, Thể thao, Thời sự có độ dài trung bình thấp hơn đáng kể (2.700–3.200 ký tự). Điều này phù hợp với thực tế: các bài phân tích sâu thường thuộc nhóm Chính trị–Kinh tế.

4 Tài liệu tham khảo

Nhóm viết bài báo cáo này không sử dụng nguồn tham khảo nào.