

PROJECT 2B:

Ứng dụng các mô hình Topic Modeling để phân tích và khám phá chủ đề trong báo Dân Trí

1st Mai Thanh Phuc, 2nd Hoang Thi Yen Nhi, 3rd Tran Trong Thanh, and Le Nhat Tung
HUTECH University, Vietnam
{Mai Thanh Phuc, Hoang Thi Yen Nhi, Tran Trong Thanh}@hutech.edu.vn, and lenhattung@hutech.edu.vn

Tóm tắt nội dung

Trong bối cảnh lượng thông tin báo chí ngày càng tăng nhanh, việc khai phá và phân tích nội dung trở thành một thách thức quan trọng nhằm hỗ trợ độc giả và nhà nghiên cứu nắm bắt xu hướng xã hội. Nghiên cứu này tập trung vào dữ liệu báo Dân Trí với 7.052 bài viết đã được tiền xử lý, áp dụng nhiều phương pháp mô hình hóa chủ đề (Topic Modeling) khác nhau bao gồm LDA (Latent Dirichlet Allocation), NMF (Non-negative Matrix Factorization), LSA (Latent Semantic Analysis) và BERTopic. Các mô hình được đánh giá bằng các chỉ số phổ biến như Coherence Score, Explained Variance và Reconstruction Error, qua đó lựa chọn số lượng chủ đề (K) tối ưu. Kết quả thực nghiệm cho thấy mỗi phương pháp có ưu và nhược điểm riêng: LDA mang lại khả năng giải thích tốt, NMF đạt coherence cao và ổn định, trong khi BERTopic tận dụng embedding ngữ nghĩa để nhóm các bài viết có tính tương đồng nội dung. Nghiên cứu góp phần cung cấp một khung tham chiếu hữu ích cho việc ứng dụng Topic Modeling vào phân tích báo chí tiếng Việt, từ đó hỗ trợ việc khám phá tri thức và quản lý thông tin hiệu quả.

Index Terms

Topic Modeling, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), BERTopic.

I. GIỚI THIỆU

Trong kỷ nguyên bùng nổ thông tin, các trang báo điện tử đóng vai trò quan trọng trong việc truyền tải tri thức và phản ánh tình hình xã hội. Tuy nhiên, khối lượng dữ liệu báo chí ngày càng lớn gây khó khăn trong việc tổng hợp, phân loại và khai thác thông tin. Đặc biệt với báo điện tử Dân Trí – một trong những nguồn tin tức phổ biến tại Việt Nam – việc tự động phát hiện và phân tích các chủ đề tiềm ẩn trong hàng nghìn bài viết có ý nghĩa thiết thực nhằm hỗ trợ độc giả nắm bắt nhanh xu hướng cũng như phục vụ công tác nghiên cứu.

Mô hình hóa chủ đề (Topic Modeling) là một trong những kỹ thuật quan trọng trong lĩnh vực Khai phá văn bản (Text Mining) và Xử lý ngôn ngữ tự nhiên (NLP). Các phương pháp kinh điển như Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) và Latent Semantic Analysis (LSA) đã được ứng dụng rộng rãi để phát hiện cấu trúc ngữ nghĩa tiềm ẩn từ dữ liệu văn bản. Gần đây, các phương pháp hiện đại như BERTopic khai thác sức mạnh của biểu diễn ngữ nghĩa (embeddings) đã mở ra hướng tiếp cận mới, mang lại hiệu quả cao hơn trong việc mô hình hóa ngữ nghĩa của văn bản.

Nghiên cứu này tập trung phân tích hơn 7.052 bài viết báo Dân Trí, đã qua tiền xử lý, nhằm so sánh hiệu quả giữa nhiều kỹ thuật mô hình hóa chủ đề. Chúng tôi triển khai và đánh giá LDA, NMF, LSA, đồng thời mở rộng với BERTopic để tận dụng sức mạnh của biểu diễn Phobert/SBERT. Các mô hình được so sánh dựa trên các chỉ số khách quan như Coherence Score, Reconstruction Error và Explained Variance, từ đó xác định số lượng chủ đề tối ưu. Kết quả phân tích cho phép đề xuất một khung tham chiếu hữu ích để áp dụng Topic Modeling vào ngữ cảnh báo chí tiếng Việt, góp phần nâng cao khả năng tổ chức và quản lý thông tin trong thực tiễn.

II. NGHIÊN CỨU LIÊN QUAN

Mô hình hóa chủ đề (Topic Modeling) đã được nghiên cứu và ứng dụng rộng rãi trong lĩnh vực khai phá văn bản (Text Mining) và xử lý ngôn ngữ tự nhiên (NLP). Trong số đó, *Latent Dirichlet Allocation (LDA)* [1] được coi là phương pháp kinh điển, dựa trên phân phối xác suất để mô hình hóa sự phân bố từ trong các tài liệu. LDA đã chứng minh hiệu quả trong nhiều bài toán phân tích văn bản như phân loại tin tức, phân tích xu hướng và tổ chức dữ liệu quy mô lớn.

Ngoài LDA, *Non-negative Matrix Factorization (NMF)* [2] là một phương pháp phân rã ma trận được ứng dụng thành công cho bài toán phân tích chủ đề. NMF cho phép tìm ra các thành phần không âm, dễ diễn giải hơn trong ngữ cảnh văn bản. Nhiều nghiên cứu đã chỉ ra rằng NMF có thể đạt được độ chặt chẽ (coherence) cao hơn so với LDA trong một số tập dữ liệu thực tế.

Latent Semantic Analysis (LSA) [3] là một hướng tiếp cận khác, khai thác kỹ thuật giảm chiều bằng *Singular Value Decomposition (SVD)* để phát hiện cấu trúc ngữ nghĩa tiềm ẩn. Mặc dù LSA có hạn chế trong việc diễn giải xác suất, nhưng nó vẫn được sử dụng như một phương pháp nền tảng để so sánh với các mô hình hiện đại.

Trong những năm gần đây, các mô hình dựa trên *embedding ngữ nghĩa* kết hợp với kỹ thuật clustering, tiêu biểu là *BERTopic* [4], đã được phát triển nhằm tận dụng khả năng biểu diễn ngữ cảnh của các mô hình ngôn ngữ tiền huấn luyện (pre-trained language models). BERTopic cho phép phát hiện các chủ đề với độ chính xác cao hơn, đặc biệt hiệu quả khi áp dụng cho dữ liệu đa ngữ như tiếng Việt.

Đối với tiếng Việt, một số nghiên cứu đã áp dụng LDA và NMF cho phân loại tin tức và phân tích dữ liệu báo chí. Các nghiên cứu cũng chỉ ra rằng việc kết hợp biểu diễn TF-IDF với mô hình phân rã ma trận hoặc embedding từ các mô hình ngôn ngữ như PhoBERT mang lại kết quả khả quan hơn. Tuy nhiên, so sánh toàn diện giữa các phương pháp truyền thống (LDA, NMF, LSA) và phương pháp hiện đại (BERTopic) trên dữ liệu báo chí tiếng Việt vẫn chưa được khai thác nhiều, từ đó mở ra khoảng trống nghiên cứu mà bài báo này hướng tới.

III. PHƯƠNG PHÁP NGHIÊN CỨU

A. Thu thập dữ liệu

Dữ liệu được thu thập từ **báo điện tử Dân Trí** — một trong những nguồn tin tức phổ biến và đa dạng chủ đề nhất tại Việt Nam. Mục tiêu là thu thập tập hợp các bài viết thuộc nhiều chuyên mục khác nhau (chính trị, kinh tế, giáo dục, sức khỏe, thể thao, giải trí, v.v.), nhằm đảm bảo dữ liệu có độ bao phủ nội dung rộng, phục vụ tốt cho quá trình phân cụm chủ đề.

1) *Công cụ và ngôn ngữ lập trình*: Việc thu thập dữ liệu được thực hiện bằng ngôn ngữ **Python**, sử dụng hai thư viện chính là:

- **Selenium**: mô phỏng thao tác người dùng trên trình duyệt, cho phép cuộn trang và tải động nội dung.
- **BeautifulSoup4**: dùng để trích xuất dữ liệu HTML (tiêu đề, ngày đăng, tác giả, nội dung) từ từng bài viết.

Ngoài ra, thư viện `pandas` được sử dụng để xử lý dữ liệu dạng bảng và lưu trữ kết quả cào dưới dạng **CSV**, giúp thuận tiện cho các bước tiền xử lý sau này.

2) *Quy trình cào dữ liệu*: Quá trình thu thập được chia làm hai giai đoạn:

a) (1) *Cào danh sách bài viết theo chuyên mục*: Trình duyệt Chrome được chạy ở chế độ `headless` để tăng tốc độ và tiết kiệm tài nguyên. Mỗi chuyên mục trên trang <https://dantri.com.vn> được tự động duyệt qua tối đa 15 trang (cuộn 10 lần mỗi trang) để thu thập các liên kết bài viết, tiêu đề, chuyên mục và thời điểm cào. Hệ thống tự động loại bỏ trùng lặp bằng cách đối chiếu với dữ liệu đã có, sau đó lưu kết quả vào tệp **CSV** kèm timestamp.

`dantri_new_unique_2025-10-12_00-50-08.csv`

b) (2) *Cào nội dung chi tiết từng bài báo*: Ở giai đoạn thứ hai, mỗi liên kết thu được từ bước trên được truy cập riêng lẻ bằng thư viện `requests`. Dữ liệu chi tiết được trích xuất gồm:

- **Tiêu đề bài viết (title)**
- **Ngày đăng (pub_date)**
- **Tác giả (author)**
- **Nội dung (content)** – gồm toàn bộ phần thân bài dưới thẻ HTML `<p>` trong vùng `singular-content`.

Để đảm bảo tính ổn định, chương trình thực hiện:

- Tạm dừng ngẫu nhiên (`time.sleep(random.uniform(0.5, 1.2))`) giữa các yêu cầu để tránh bị chặn IP.
- Bỏ qua các bài không tải được hoặc thiếu nội dung chính.
- Gộp dữ liệu đã cào mới với danh sách link cũ bằng `pandas.merge()` theo khóa “link”.

Sau khi hoàn tất, toàn bộ dữ liệu được lưu lại trong một tệp **CSV** tổng hợp, ví dụ:

`dantri_crawl_full_2025-10-12_01-40-30.csv`

c) (3) *Quy mô dữ liệu*: Tổng cộng có **7.580 bài viết** được thu thập, trải rộng trên hơn 20 chuyên mục, bao gồm cả các chủ đề phổ biến như thời sự, thể thao, sức khỏe, giải trí và kinh doanh. Mỗi bản ghi trong dữ liệu gồm các trường:

`[category, title, link, pub_date, author, content, crawl_date]`

3) *Đánh giá chất lượng dữ liệu*: Sau khi cào xong, dữ liệu được kiểm tra và làm sạch sơ bộ nhằm loại bỏ:

- Các bài viết trùng lặp hoặc bị lỗi khi tải nội dung.
- Những bài có phần nội dung quá ngắn (dưới 30 ký tự).
- Các bài không có tiêu đề hoặc ngày đăng.

Quy trình này giúp đảm bảo dữ liệu đầu vào có chất lượng tốt, phục vụ hiệu quả cho các bước tiền xử lý ngôn ngữ và biểu diễn đặc trưng tiếp theo.

B. Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu thô từ báo điện tử Dân Trí, bước tiếp theo là tiến hành **tiền xử lý dữ liệu văn bản** nhằm loại bỏ nhiễu, chuẩn hóa ngôn ngữ và chuyển đổi dữ liệu thành dạng có thể sử dụng trong mô hình học máy. Toàn bộ quá trình được thực hiện bằng ngôn ngữ **Python**, sử dụng các thư viện chính như `pandas`, `re`, `underthesea`, `datasketch`, `scikit-learn`, và `matplotlib`.

1) *Đọc và lựa chọn dữ liệu đầu vào*: Từ tệp dữ liệu gốc `dantri_crawl_full.csv`, ba trường quan trọng được giữ lại:

- **category**: chuyên mục của bài viết.
- **content**: nội dung chi tiết của bài báo.
- **clean_text**: văn bản đã được làm sạch sơ bộ (lowercase, bỏ ký tự đặc biệt).

2) *Làm sạch và chuẩn hóa văn bản*: Các bước làm sạch được thực hiện bao gồm:

- **Chuyển chữ thường toàn bộ** để giảm sự phân biệt giữa chữ hoa và chữ thường.
- **Loại bỏ liên kết URL, ký tự đặc biệt, số, emoji và dấu câu** bằng biểu thức chính quy.
- **Chuẩn hóa khoảng trắng** nhằm loại bỏ các dấu cách thừa.
- **Tách từ (word tokenization)** bằng thư viện `Underthesea`, đảm bảo xử lý đúng các cụm từ ghép tiếng Việt.

3) *Loại bỏ trùng lặp và bản ghi rỗng*: Hai cấp độ loại bỏ trùng lặp được áp dụng:

- **Trùng lặp tuyệt đối**: loại bỏ bằng `drop_duplicates()` dựa trên cột `clean_text`.
- **Trùng lặp gần (near-duplicate)**: áp dụng `MinHash + Locality Sensitive Hashing (LSH)` từ thư viện `datasketch` để phát hiện và loại bỏ các bài viết có độ tương đồng lớn hơn 90%.

Ngoài ra, các văn bản trống (`clean_text = ""`) cũng được loại bỏ.

4) *Loại bỏ văn bản ngắn*: Các bài viết quá ngắn sẽ không cung cấp đủ ngữ cảnh để mô hình học được phân bố từ vựng. Do đó:

- Loại bỏ các văn bản có ít hơn **100 tokens**.

Bước này giúp dữ liệu ổn định hơn cho các mô hình như LDA, NMF hoặc LSA.

5) *Phân tích và xử lý mất cân bằng dữ liệu*: Sau khi lọc, dữ liệu còn lại **7.052 bài viết**. Tuy nhiên, phân bố các chuyên mục (*category*) vẫn không đồng đều, một số chuyên mục rất ít bài. Để giải quyết:

- Các chuyên mục có số lượng cực ít (dưới 50 bài) được gộp vào nhóm **“other”**.
- Với các chuyên mục còn lại, tiến hành thống kê và lựa chọn kỹ thuật xử lý mất cân bằng (nếu cần) như oversampling hoặc weighting trong huấn luyện.

6) *Trích xuất từ khóa đặc trưng cho từng chuyên mục*: Để hỗ trợ việc kiểm tra chất lượng dữ liệu và định hướng gộp nhóm, TF-IDF được áp dụng trên từng *category* để lấy ra **top-words**. Điều này giúp xác định các chủ đề chính và các chuyên mục dễ bị chồng lấn.

7) *Gộp chuyên mục*: Dựa trên phân tích top-words, hai chuyên mục hiếm *Infographic* và *Xã hội* được gộp lẫn lượt vào *Ô tô – Xe máy* và *Thời sự*. Việc này đảm bảo phân bố dữ liệu cân bằng hơn và hạn chế ảnh hưởng tiêu cực tới kết quả mô hình hóa chủ đề.

8) *Lưu kết quả tiền xử lý*: Dữ liệu cuối cùng được lưu thành nhiều phiên bản phục vụ các bước tiếp theo:

- `dantri_qc_pass.csv`: dữ liệu đã qua QC (loại trùng lặp, bài ngắn).
- `dantri_qc_grouped.csv`: dữ liệu sau khi gộp chuyên mục.

C. Biểu diễn đặc trưng văn bản

Để các mô hình chủ đề có thể xử lý văn bản, dữ liệu thô cần được ánh xạ thành **vector số** biểu diễn ý nghĩa và cấu trúc ngữ nghĩa. Trong nghiên cứu này, ba hướng biểu diễn đặc trưng được sử dụng:

1) *TF-IDF cho NMF và LSA*: Đối với hai phương pháp phân rã ma trận **NMF (Non-negative Matrix Factorization)** và **LSA (Latent Semantic Analysis)**, đầu vào là ma trận **TF-IDF (Term Frequency – Inverse Document Frequency)**:

- Số chiều đặc trưng được giới hạn ở mức 50.000 từ phổ biến nhất.
- Sử dụng cả đơn từ (unigram) và cụm hai từ (bigram) với `ngram_range = (1, 2)`.
- Lọc bỏ từ xuất hiện trong ít hơn 5 văn bản hoặc nhiều hơn 85% văn bản (`min_df = 5, max_df = 0.85`).

Ma trận TF-IDF sau khi tính toán được lưu lại (`tfidf_matrix.pkl`) để tái sử dụng nhiều lần mà không cần tính toán lại.

2) *Bag-of-Words cho LDA*: Khác với NMF và LSA, mô hình **LDA (Latent Dirichlet Allocation)** là một mô hình sinh xác suất, yêu cầu đầu vào là **ma trận đếm (Bag-of-Words)** thay vì TF-IDF. Do đó, dữ liệu được vector hóa bằng `CountVectorizer` với các tham số:

- `min_df = 5`: loại bỏ các từ xuất hiện trong quá ít văn bản.
- `max_df = 0.9`: loại bỏ các từ xuất hiện quá phổ biến (trên 90% văn bản).

Điều này giúp mô hình tránh hiện tượng **“chủ đề rác”**, vốn xảy ra khi áp dụng TF-IDF cho LDA.

3) *Embedding ngữ cảnh bằng PhoBERT (SBERT)*: Để tận dụng sức mạnh của các mô hình ngôn ngữ hiện đại, nghiên cứu sử dụng **PhoBERT** (thông qua mô hình `keepitreal/vietnamese-sbert` trong thư viện `SentenceTransformers`).

- Mỗi văn bản được mã hóa thành vector kích thước cố định (768 chiều).
- Quá trình mã hóa được tăng tốc bằng GPU (CUDA) khi khả dụng.
- Embedding tạo ra có khả năng giữ ngữ nghĩa ngữ cảnh, vượt trội so với TF-IDF hoặc BoW vốn chỉ dựa trên tần suất.
- Tất cả vector embedding được lưu dưới dạng tệp NumPy (`.npy`) để phục vụ các bước phân cụm hoặc trực quan hóa.

4) *Tổng kết*: Như vậy, ba dạng biểu diễn đã được áp dụng:

- **TF-IDF** cho NMF và LSA.
- **Bag-of-Words** cho LDA.
- **PhoBERT Embedding** cho hướng tiếp cận hiện đại dựa trên ngữ cảnh.

Việc kết hợp nhiều kỹ thuật biểu diễn cho phép so sánh toàn diện giữa các phương pháp truyền thống và các mô hình ngôn ngữ hiện đại.

D. Huấn luyện mô hình

Sau khi dữ liệu văn bản được chuẩn hoá và biểu diễn dưới dạng đặc trưng số, bước tiếp theo là **huấn luyện các mô hình trích xuất chủ đề**. Trong nghiên cứu này, chúng tôi triển khai bốn phương pháp tiêu biểu: Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), và BERTopic dựa trên PhoBERT. Các mô hình này được lựa chọn vì đại diện cho cả hướng tiếp cận truyền thống (dựa trên ma trận đếm hoặc TF-IDF) và hiện đại (dựa trên embedding ngữ cảnh).

1. Latent Dirichlet Allocation (LDA)

LDA là một mô hình sinh xác suất (probabilistic generative model) được Blei et al. (2003) đề xuất. Ý tưởng chính là mỗi văn bản được coi như một phân phối xác suất trên các chủ đề tiềm ẩn, và mỗi chủ đề lại là một phân phối xác suất trên các từ vựng. Quá trình sinh văn bản có thể mô tả như sau:

- Chọn phân phối chủ đề $\theta_d \sim \text{Dirichlet}(\alpha)$ cho văn bản d .
- Với mỗi từ w trong văn bản:
 - Chọn một chủ đề $z \sim \text{Multinomial}(\theta_d)$.
 - Sinh từ w từ phân phối $\phi_z \sim \text{Dirichlet}(\beta)$.

Trong nghiên cứu, LDA được huấn luyện trên ma trận Bag-of-Words với nhiều giá trị số chủ đề $k \in [8, 20]$. Chất lượng mô hình được đánh giá thông qua chỉ số **Coherence (c_v)** nhằm phản ánh mức độ gắn kết ngữ nghĩa giữa các từ trong một chủ đề.

2. Non-negative Matrix Factorization (NMF)

NMF là phương pháp phân rã ma trận không âm (Lee Seung, 1999). Ý tưởng là ma trận TF-IDF $X \in \mathbb{R}^{m \times n}$ (với m văn bản, n đặc trưng) được xấp xỉ bởi tích của hai ma trận không âm $W \in \mathbb{R}^{m \times k}$ và $H \in \mathbb{R}^{k \times n}$:

$$X \approx W \cdot H$$

Trong đó:

- W : biểu diễn phân phối chủ đề của mỗi văn bản.
- H : biểu diễn phân phối từ của mỗi chủ đề.

NMF đảm bảo tính diễn giải dễ dàng do ràng buộc không âm, mỗi chủ đề là một tổ hợp tuyến tính của các từ. Chúng tôi đánh giá mô hình bằng **Coherence (c_v)** và **Reconstruction Error**:

$$\text{Error} = \|X - WH\|_F$$

Trong đó $\|\cdot\|_F$ là chuẩn Frobenius.

3. Latent Semantic Analysis (LSA)

LSA (Deerwester et al., 1990) là phương pháp giảm chiều dựa trên phân rã giá trị kỳ dị (SVD). Cho ma trận TF-IDF X , phân rã:

$$X \approx U \Sigma V^T$$

Trong đó:

- U : ma trận biểu diễn văn bản trong không gian tiềm ẩn.
- V : ma trận biểu diễn từ trong không gian tiềm ẩn.
- Σ : ma trận đường chéo chứa các giá trị kỳ dị, phản ánh độ quan trọng của từng chủ đề.

LSA nắm bắt được cấu trúc ngữ nghĩa tiềm ẩn bằng cách giảm chiều và loại bỏ nhiễu. Mô hình được đánh giá bằng **Coherence (c_v)** và **Explained Variance Ratio**, cho biết tỷ lệ phương sai dữ liệu được giữ lại sau giảm chiều.

4. PhoBERT + BERTopic

BERTopic (Grootendorst, 2020) là một phương pháp hiện đại dựa trên **embedding ngữ cảnh**. Trong nghiên cứu này, chúng tôi sử dụng **PhoBERT** (Nguyen, 2020) – một mô hình Transformer tối ưu cho tiếng Việt – để tạo vector embedding cho mỗi văn bản. Quy trình BERTopic bao gồm:

- 1) Biểu diễn văn bản bằng PhoBERT (hoặc SBERT phiên bản tiếng Việt).
- 2) Giảm chiều embedding bằng **UMAP**.
- 3) Gom cụm các văn bản bằng **HDBSCAN**.
- 4) Trích xuất từ khoá đặc trưng cho mỗi cụm bằng **c-TF-IDF**.

Ưu điểm lớn nhất của BERTopic là **không cần chỉ định trước số chủ đề**, đồng thời cho phép gợi ý nhãn chủ đề rõ ràng hơn thông qua các từ khóa tự động.

1) *Chiến lược lựa chọn mô hình*: Đối với từng mô hình, chúng tôi huấn luyện với nhiều giá trị k , sau đó so sánh các chỉ số đánh giá (Coherence, Reconstruction Error, Explained Variance) và trực quan hóa bằng biểu đồ **elbow**. Mô hình cuối cùng được chọn dựa trên:

- Coherence cao nhất và ổn định.
- Reconstruction Error/Explained Variance hợp lý (không quá cao hoặc thấp).
- Phân bố văn bản trên các chủ đề cân bằng.

E. Chỉ số đánh giá

Để đánh giá chất lượng của các mô hình phân tích chủ đề (topic modeling), luận văn sử dụng hai nhóm chỉ số chính: **độ đo nội tại** (intrinsic metrics) và **độ đo ngoại tại** (extrinsic metrics). Trong phạm vi này, tập trung vào năm chỉ số phổ biến: *Coherence Score*, *Perplexity*, *Reconstruction Error*, *Adjusted Rand Index (ARI)* và *Normalized Mutual Information (NMI)*.

1) *Topic Coherence*: Chỉ số **Coherence** đo lường mức độ gắn kết ngữ nghĩa giữa các từ quan trọng trong cùng một chủ đề. Một trong những biến thể phổ biến là C_v , được tính dựa trên điểm đồng xuất hiện của các từ trong cửa sổ ngữ cảnh:

$$C_v = \frac{1}{|W|} \sum_{w_i, w_j \in W} \text{NPMI}(w_i, w_j)$$

Trong đó:

- W là tập từ khóa của một chủ đề.
- $\text{NPMI}(w_i, w_j)$ là chỉ số Normalized Pointwise Mutual Information giữa hai từ w_i, w_j .

Chỉ số C_v càng cao chứng tỏ chủ đề càng dễ hiểu và có ý nghĩa [5].

2) *Perplexity*: Chỉ số **Perplexity** thường được sử dụng trong LDA để đo mức độ phù hợp của mô hình với dữ liệu chưa thấy. Perplexity càng thấp nghĩa là mô hình dự đoán phân bố từ trong văn bản càng tốt:

$$\text{Perplexity}(D) = \exp \left\{ -\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\}$$

Trong đó:

- M là số văn bản.
- w_d là tập từ trong văn bản d .
- N_d là số từ trong văn bản d .

3) *Reconstruction Error*: Đối với các mô hình dựa trên phân rã ma trận như NMF, chỉ số **Reconstruction Error** được sử dụng để đo mức độ sai khác giữa ma trận gốc X và ma trận tái tạo $\hat{X} = W \cdot H$:

$$E = \|X - WH\|_F^2 = \sum_{i,j} (X_{ij} - (WH)_{ij})^2$$

Trong đó:

- X là ma trận đặc trưng đầu vào (TF-IDF).
- W, H là hai ma trận nhân tố không âm trong NMF.
- $\|\cdot\|_F$ ký hiệu chuẩn Frobenius.

Giá trị E càng nhỏ thì mô hình tái tạo dữ liệu gốc càng chính xác.

4) *Adjusted Rand Index (ARI)*: Chỉ số **ARI** đo lường mức độ tương đồng giữa hai phân cụm (clusterings): nhãn dự đoán từ mô hình và nhãn gốc (ground-truth). ARI hiệu chỉnh Rand Index để loại bỏ sự ngẫu nhiên:

$$ARI = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]}$$

Trong đó RI (Rand Index) là tỷ lệ cặp điểm dữ liệu được phân cụm đúng (cùng cụm hoặc khác cụm). Giá trị ARI nằm trong khoảng $[-1, 1]$, với 1 là trùng khớp hoàn hảo, 0 là ngẫu nhiên [6].

5) *Normalized Mutual Information (NMI)*: Chỉ số **NMI** đo mức độ chia sẻ thông tin giữa nhãn gốc và nhãn dự đoán, chuẩn hóa trong khoảng $[0, 1]$:

$$NMI(U, V) = \frac{2 \cdot I(U; V)}{H(U) + H(V)}$$

Trong đó:

- $I(U; V)$ là thông tin tương hỗ giữa phân cụm U và V .
- $H(U), H(V)$ là entropy của hai phân cụm.

Giá trị NMI càng cao thì mô hình phân cụm càng gần với nhãn thực tế [7].

IV. THIẾT LẬP THÍ NGHIỆM

A. Môi trường thực nghiệm

Toàn bộ quá trình thực nghiệm được thực hiện trên máy tính cá nhân với cấu hình như sau:

Bảng I
CẤU HÌNH MÔI TRƯỜNG THỰC NGHIỆM

Thành phần	Thông tin
Hệ điều hành	Windows 11 64-bit
Bộ xử lý (CPU)	Intel Core i7-12700H (14 nhân, 20 luồng)
RAM	16 GB DDR5
Ngôn ngữ lập trình	Python 3.13.2
IDE	PyCharm Community Edition 2022.2.3

Thư viện chính sử dụng:

- `pandas`, `numpy`: Xử lý và quản lý dữ liệu dạng bảng.
- `re`: Biểu thức chính quy để làm sạch văn bản.
- `underthesea`: Tách từ tiếng Việt và chuẩn hóa ngôn ngữ.
- `datasketch`: MinHash và LSH để phát hiện văn bản trùng lặp hoặc gần trùng.
- `scikit-learn`: Cung cấp các bộ biến đổi văn bản (TF-IDF, CountVectorizer), các thuật toán phân rã ma trận (NMF, TruncatedSVD) và LDA.
- `gensim`: Tính toán chỉ số coherence và hỗ trợ các mô hình topic modeling.
- `imbalanced-learn (imblearn)`: Kỹ thuật oversampling để cân bằng dữ liệu.
- `matplotlib`, `seaborn`, `wordcloud`: Trực quan hóa kết quả (biểu đồ, WordCloud).
- `sentence-transformers (PhoBERT/SBERT)`: Sinh embedding ngữ nghĩa cho văn bản.
- `joblib`, `pickle`: Lưu và tải lại mô hình cũng như ma trận đặc trưng.

B. Dữ liệu đầu vào

Tập dữ liệu được sử dụng trong nghiên cứu này được thu thập từ **báo điện tử Dân Trí** thông qua kỹ thuật *web scraping*. Quá trình thu thập diễn ra trong nhiều chuyên mục khác nhau như: *Thời sự*, *Kinh doanh*, *Giáo dục*, *Giải trí*, *Thể thao*, *Sức khỏe*, *Tình yêu – Giới tính*, *Pháp luật*, *Ô tô – Xe máy*, v.v.

Sau khi loại bỏ dữ liệu nhiễu và xử lý trùng lặp, tập dữ liệu cuối cùng bao gồm:

- **Số lượng văn bản**: 7.052 bài báo hợp lệ sau bước tiền xử lý.
- **Trường dữ liệu**:
 - `category` – chuyên mục của bài viết.
 - `content` – nội dung chi tiết bài báo.
 - `clean_text` – văn bản đã được chuẩn hóa (làm sạch, tách từ, loại bỏ từ dừng).
- **Đặc điểm**: Các văn bản có độ dài trung bình khoảng 250–500 từ, đảm bảo ngữ cảnh đủ lớn cho các mô hình khai phá chủ đề.

Ngoài ra, để tránh hiện tượng mất cân bằng giữa các lớp, dữ liệu đã được **cân bằng lại** bằng kỹ thuật Oversampling, từ đó tạo ra phân bố đồng đều hơn giữa các chuyên mục. Tập dữ liệu cuối cùng được lưu dưới dạng CSV và sử dụng làm đầu vào cho các bước *biểu diễn đặc trưng văn bản (embedding)* và *huấn luyện mô hình chủ đề (topic modeling)*.

Bảng II
PHÂN BỐ SỐ LƯỢNG BÀI VIẾT THEO CHUYÊN MỤC GỐC (CATEGORY)

Chuyên mục	Số bài viết
Giáo dục	625
Thời sự	608
Giải trí	575
Sức khỏe	559
Thể thao	506
Kinh doanh	501
Ô tô - Xe máy	491
Việc làm	472
Du lịch	470
Bạn đọc	421
Tình yêu - Giới tính	412
Thể giới	388
Pháp luật	385
Tâm điểm	253
Văn hóa	199
Nhịp sống trẻ	187

C. Thuật toán và quy trình thực hiện

Trong nghiên cứu này, chúng tôi áp dụng nhiều thuật toán **khai phá chủ đề (topic modeling)** khác nhau nhằm so sánh và đánh giá hiệu quả giữa các phương pháp truyền thống và phương pháp hiện đại. Các thuật toán chính bao gồm:

- **Latent Dirichlet Allocation (LDA)** [1]: Đây là mô hình sinh xác suất, giả định mỗi tài liệu là một phân phối Dirichlet trên các chủ đề và mỗi chủ đề là một phân phối Dirichlet trên các từ. LDA hoạt động hiệu quả trên ma trận **Bag-of-Words** hoặc **CountVectorizer**.
- **Non-negative Matrix Factorization (NMF)** [2]: NMF phân rã ma trận TF-IDF thành hai ma trận con không âm, cho phép diễn giải các chủ đề dựa trên trọng số của từ khóa. NMF thường cho ra kết quả dễ giải thích hơn so với LDA.
- **Latent Semantic Analysis (LSA)** [3]: LSA sử dụng **Truncated SVD** để giảm chiều không gian của ma trận TF-IDF, từ đó tìm ra các thành phần tiềm ẩn (latent components) phản ánh mối quan hệ ngữ nghĩa giữa các từ và tài liệu.
- **BERTopic** [4]: Là phương pháp hiện đại kết hợp **Sentence-BERT embeddings** với các thuật toán giảm chiều (*UMAP*) và phân cụm (*HDBSCAN*). BERTopic không dựa vào ma trận TF-IDF mà khai thác sức mạnh của biểu diễn ngữ nghĩa từ mô hình Transformer, cho phép phát hiện chủ đề có chất lượng cao hơn trên dữ liệu tiếng Việt.

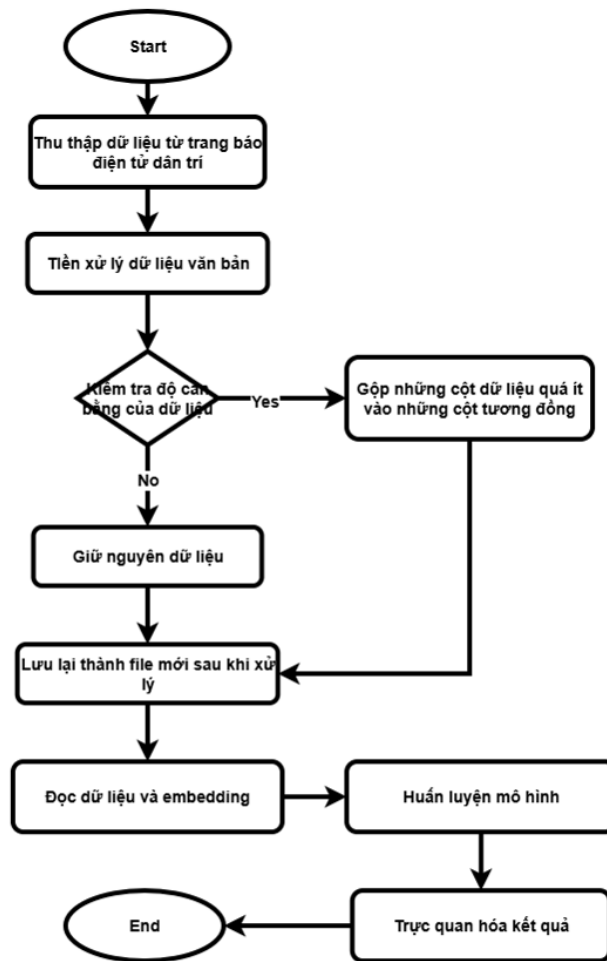
Quy trình thực hiện được mô tả qua các bước sau:

- 1) **Tiền xử lý dữ liệu**: Làm sạch văn bản, chuẩn hóa, tách từ, loại bỏ stopwords, loại bỏ trùng lặp và cân bằng lại dữ liệu giữa các chuyên mục.
- 2) **Biểu diễn đặc trưng văn bản**:
 - TF-IDF cho NMF và LSA.
 - CountVectorizer cho LDA.
 - Sentence-BERT embeddings cho BERTopic.
- 3) **Huấn luyện mô hình chủ đề**: Mỗi mô hình được huấn luyện với nhiều giá trị số chủ đề K (từ 8 đến 20) nhằm đánh giá và lựa chọn K tối ưu.
- 4) **Đánh giá mô hình**: Các chỉ số chính được sử dụng là **Coherence score (C_v)**, **Reconstruction error** (đối với NMF), và **Explained Variance** (đối với LSA), Ari, Nmi.
- 5) **Trực quan hóa và gán nhãn chủ đề**: Kết quả được trực quan hóa bằng **WordCloud**, biểu đồ cột và biểu đồ phân bố. Các chủ đề được gán nhãn dựa trên tập từ khóa nổi bật (top words) và các phương pháp hỗ trợ như *KeyBERT*.

Quy trình trên đảm bảo tính so sánh toàn diện giữa các mô hình truyền thống và hiện đại, đồng thời cung cấp kết quả trực quan, dễ diễn giải cho dữ liệu báo chí tiếng Việt.

D. Mô tả quy trình thực hiện

Quy trình nghiên cứu và triển khai được thể hiện ngắn gọn qua sơ đồ khối ở Hình 1. Sơ đồ minh họa tuần tự các bước từ thu thập dữ liệu, tiền xử lý, biểu diễn đặc trưng, huấn luyện mô hình chủ đề cho đến đánh giá và trực quan hóa kết quả.



Hình 1. Quy trình tổng quan của hệ thống phân tích chủ đề.

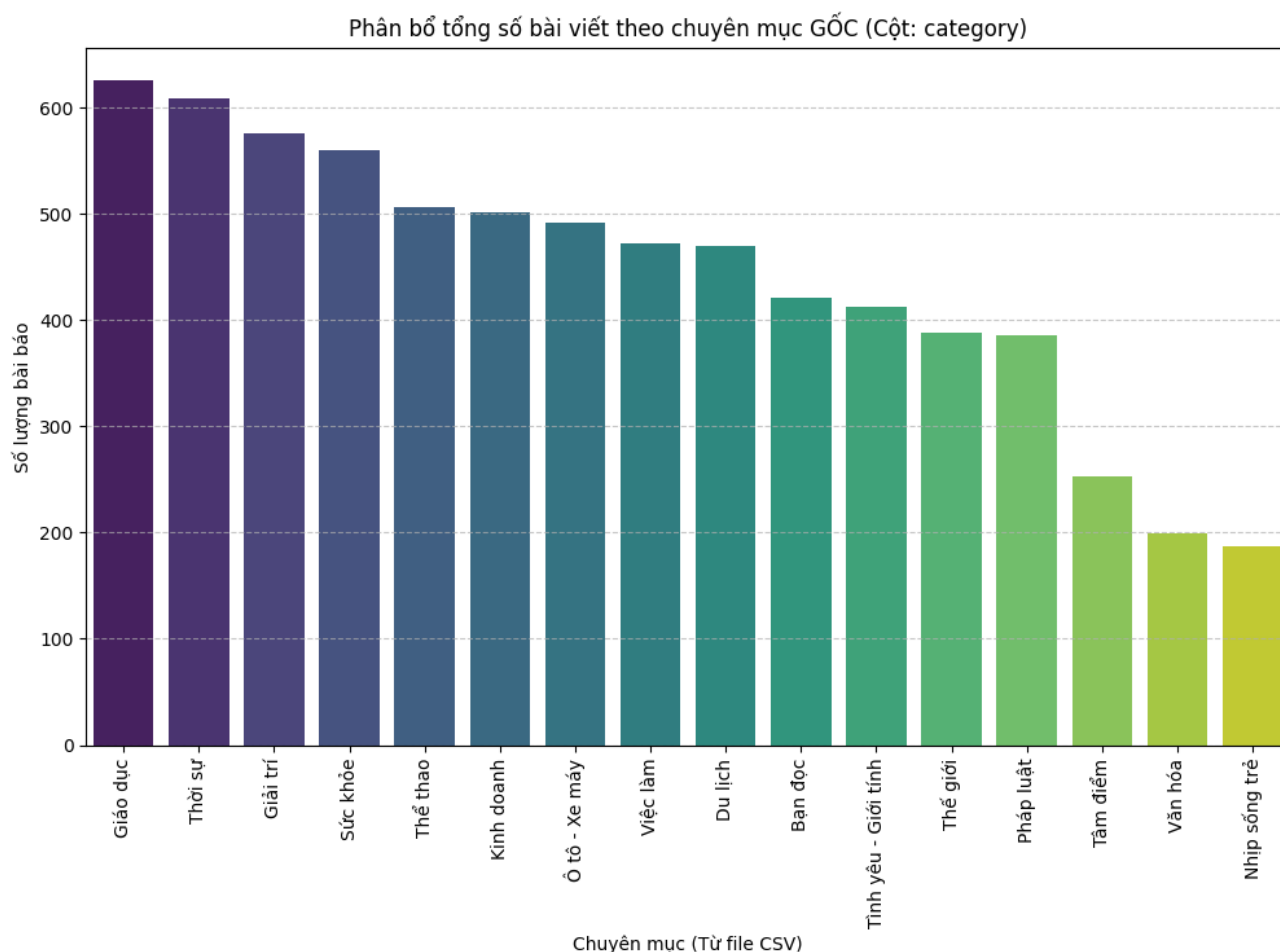
V. KẾT QUẢ VÀ THẢO LUẬN

A. Tổng quan dữ liệu

Sau quá trình thu thập và tiền xử lý, tập dữ liệu cuối cùng bao gồm **7.052 bài báo** từ báo điện tử Dân Trí, thuộc **16 chuyên mục chính**. Mỗi chuyên mục đại diện cho một mảng thông tin khác nhau như *Giáo dục*, *Thời sự*, *Giải trí*, *Sức khỏe*, *Kinh doanh*, *Thể thao*, v.v. Tuy nhiên, số lượng bài viết giữa các chuyên mục không đồng đều, phản ánh đặc trưng thực tế khi một số chuyên mục có nhiều bài báo hơn hẳn.

Hình 2 thể hiện phân bố tổng số bài viết theo từng chuyên mục gốc. Ta có thể nhận thấy rằng:

- Các chuyên mục như **Giáo dục**, **Thời sự**, **Giải trí**, **Sức khỏe** và **Thể thao** chiếm tỷ lệ lớn trong tập dữ liệu (trên 500 bài viết mỗi chuyên mục).
- Một số chuyên mục ít phổ biến hơn, ví dụ **Văn hóa**, **Nhịp sống trẻ**, **Tâm điểm**, chỉ có dưới 250 bài viết.
- Sự mất cân bằng này đặt ra thách thức cho việc huấn luyện mô hình, khi mô hình có xu hướng học tốt ở các lớp nhiều dữ liệu nhưng khó phân biệt các lớp ít dữ liệu.



Hình 2. Phân bố số lượng bài viết theo chuyên mục gốc trong tập dữ liệu.

Nhìn chung, tập dữ liệu có quy mô đủ lớn để tiến hành các phương pháp *topic modeling*, đồng thời cũng phản ánh thách thức về **mất cân bằng dữ liệu**, cần cân nhắc trong quá trình huấn luyện và đánh giá mô hình.

B. Biểu diễn đặc trưng văn bản (Embedding)

Sau khi dữ liệu đã được tiền xử lý, bước tiếp theo là chuyển đổi văn bản thành các vector số để phục vụ cho các mô hình khai phá chủ đề. Việc biểu diễn đặc trưng văn bản đóng vai trò quan trọng vì chất lượng embedding ảnh hưởng trực tiếp đến hiệu quả của các thuật toán học máy và mô hình chủ đề. Trong nghiên cứu này, ba phương pháp chính được sử dụng:

1) *TF-IDF (Term Frequency - Inverse Document Frequency)*: Phương pháp **TF-IDF** biểu diễn văn bản dưới dạng ma trận thưa, trong đó mỗi phần tử thể hiện mức độ quan trọng của một từ trong một tài liệu so với toàn bộ tập dữ liệu.

2) *Bag-of-Words (BoW)*: Đối với mô hình **LDA**, dữ liệu không được biểu diễn bằng TF-IDF mà sử dụng **Bag-of-Words**. Phương pháp này chỉ đếm tần suất xuất hiện của từ trong mỗi văn bản và không gán trọng số. Việc này phù hợp với giả định phân phối xác suất của LDA, trong đó mỗi văn bản là một phân phối trên các chủ đề, và mỗi chủ đề là một phân phối trên các từ.

3) *Sentence-BERT (SBERT) / PhoBERT*: Để khai thác ngữ nghĩa ngữ cảnh trong tiếng Việt, nghiên cứu sử dụng mô hình **PhoBERT-SBERT** đã được huấn luyện sẵn. Các mô hình dạng Transformer như PhoBERT tạo embedding ở mức câu/văn bản, nhờ đó biểu diễn được mối quan hệ ngữ nghĩa giữa các từ trong cùng ngữ cảnh. Quá trình mã hoá được thực hiện bằng thư viện `sentence-transformers`, với đầu ra là vector chiều 768. Embedding này được dùng làm đầu vào cho mô hình **BERTopic**.

4) Kết quả đầu ra:

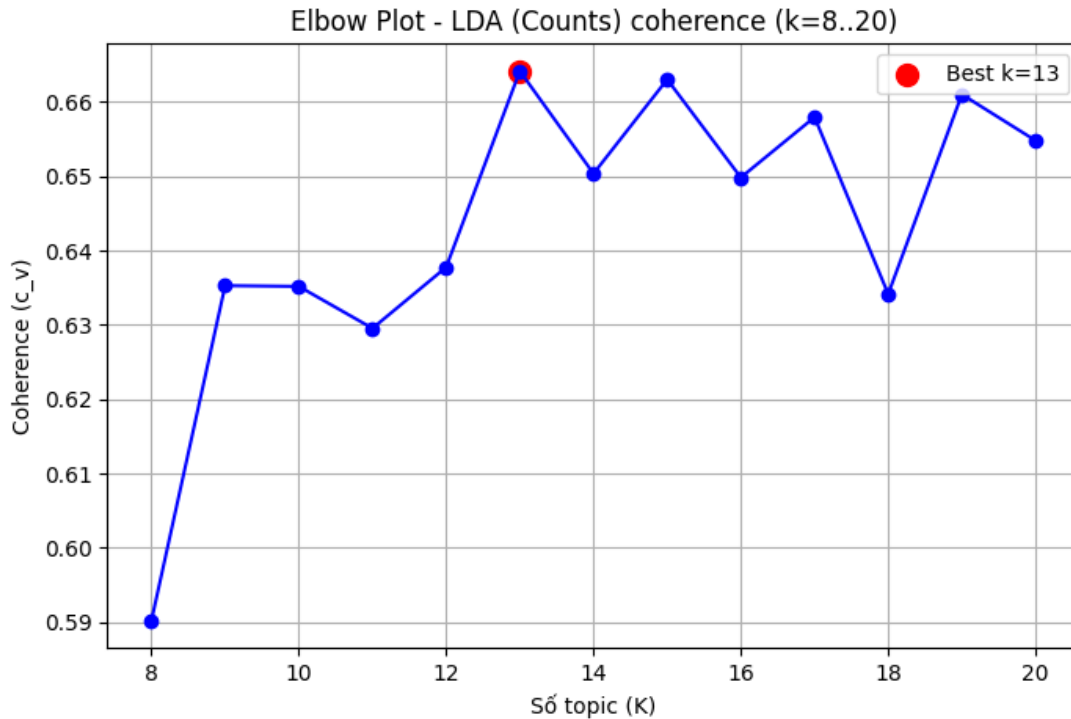
- TF-IDF: ma trận (7052, 50000) đặc trưng, dùng cho NMF và LSA.
- BoW: ma trận đếm (7052, 16255) là số lượng từ vệtng sau khi lọc, dùng cho LDA.
- PhoBERT-SBERT: ma trận embedding (7052, 768), dùng cho BERTopic.

Việc kết hợp nhiều phương pháp embedding cho phép so sánh hiệu quả của các mô hình truyền thống (LDA, NMF, LSA) và mô hình hiện đại (BERTopic).

C. Huấn luyện mô hình

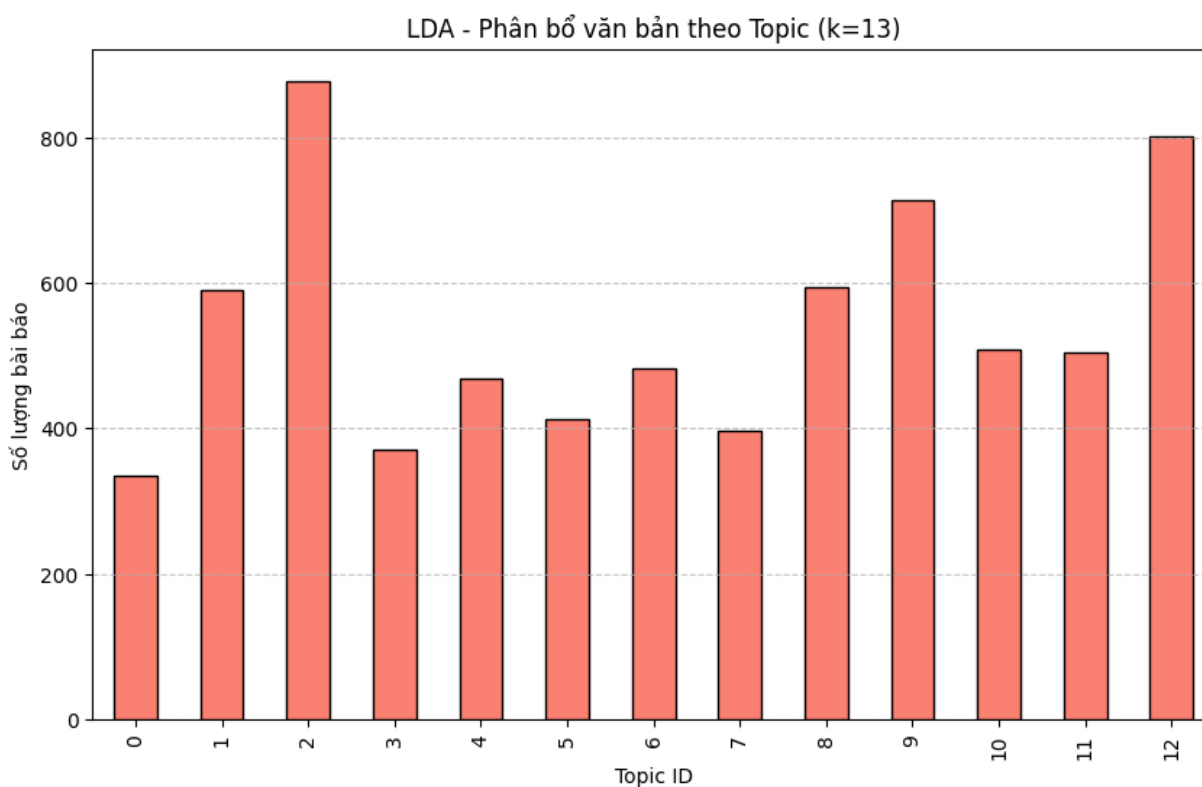
1. *Latent Dirichlet Allocation (LDA)*: Trong nghiên cứu này, thuật toán **Latent Dirichlet Allocation (LDA)** được áp dụng trên ma trận đếm từ vựng (*Bag-of-Words*) thay vì TF-IDF, nhằm đảm bảo tính chất xác suất của mô hình. Số lượng chủ đề (K) được lựa chọn dựa trên chỉ số *Coherence Score* với giá trị từ $K = 8$ đến $K = 20$.

Kết quả (Hình 3) cho thấy mô hình đạt coherence cao nhất tại $K = 13$ với $c_v = 0.6642$, đây là giá trị tối ưu để huấn luyện mô hình cuối cùng.



Hình 3. Elbow Plot của chỉ số Coherence theo số lượng topic (LDA, Bag-of-Words)

Sau khi huấn luyện mô hình LDA với $K = 13$, phân bố số lượng văn bản trên từng topic cho thấy sự đồng đều tương đối, không còn tình trạng dồn dữ liệu vào một nhóm duy nhất (Hình 4).



Hình 4. Phân bố văn bản theo 13 topic LDA

Các chủ đề chính trích xuất từ LDA ($K = 13$):

- **Topic 00:** nga, ukraine, tổng_thống, mỹ, trump, có_thể, tấn_công, bay, khu_vực, lực_lượng, máy_bay, tên_lửa, quân_đội, uav, ảnh
- **Topic 01:** việt_nam, phát_triển, doanh_nghiệp, công_nghệ, kinh_tế, có_thể, quốc_tế, không_chỉ, quốc_gia, đầu_tư, xây_dựng, bền_vững, hệ_thống, giúp, thị_trường
- **Topic 02:** công_an, xe, hành_vì, quy_định, đường, đi, vụ, vi_phạm, xử_lý, tài_xế, giao_thông, trường_hợp, có_thể, cơ QUAN, điều_tra
- **Topic 03:** việt_nam, tỉnh, tổ_chức, chủ_tịch, phát_triển, văn_hóa, ban, xây_dựng, nhân_dân, chính_trị, công_tác, đảng, ủy, thực_hiện, đất_nước
- **Topic 04:** giải, đấu, trận, việt_nam, đội_tuyển, thi_đấu, đội, giành, cầu_thủ, malaysia, vòng, ảnh, hai, vdv, chiến_thắng
- **Topic 05:** xe, đồng, mẫu, triệu, giá, ảnh, có_thể, hãng, khách_hàng, thiết_kế, sử_dụng, hai, đi, việt_nam, ô_tô
- **Topic 06:** hàng, du_khách, ảnh, quán, món, đồng, du_lịch, sản_phẩm, đi, gia_đình, hà_nội, khu, người_dân, trải_nghiệm, chia_sẻ
- **Topic 07:** tỉnh, khu_vực, mưa, người_dân, sông, xã, bão, biển, ảnh, ngập, khu, lũ, đường, cầu, tàu
- **Topic 08:** đồng, tỷ, công_ty, tiền, lao_động, triệu, quy_định, bị_cáo, giá, thuế, usd, tài_sản, vàng, khoản, đối_với
- **Topic 09:** đi, chồng, mẹ, hai, vợ, gia_đình, có_thể, gái, sống, câu_chuyện, tiền, trai, chia_sẻ, trẻ, công_việc
- **Topic 10:** có_thể, bệnh, bệnh_viện, bác_sĩ, ung_thư, điều_trị, bệnh_nhân, y_tế, giúp, thuốc, nguy_cơ, khỏe, tình_trạng, phẫu_thuật, cơ_thể
- **Topic 11:** trường, đại_học, học_sinh, học, giáo_dục, thi, sinh_viên, đào_tạo, giáo_viên, ngành, học_tập, nhà_trường, lớp, tốt_nghiệp, thí_sinh
- **Topic 12:** phim, ảnh, khán_giả, chia_sẻ, diễn_viên, nữ, diễn, nhân_vật, việt_nam, nghệ_thuật, tham_gia, trẻ, đẹp, ca_sĩ, nam

Kết quả này khẳng định rằng LDA có khả năng phân tách các nhóm chủ đề rõ ràng và mang tính diễn giải cao, phù hợp với tập dữ liệu báo chí tiếng Việt.

2. NMF

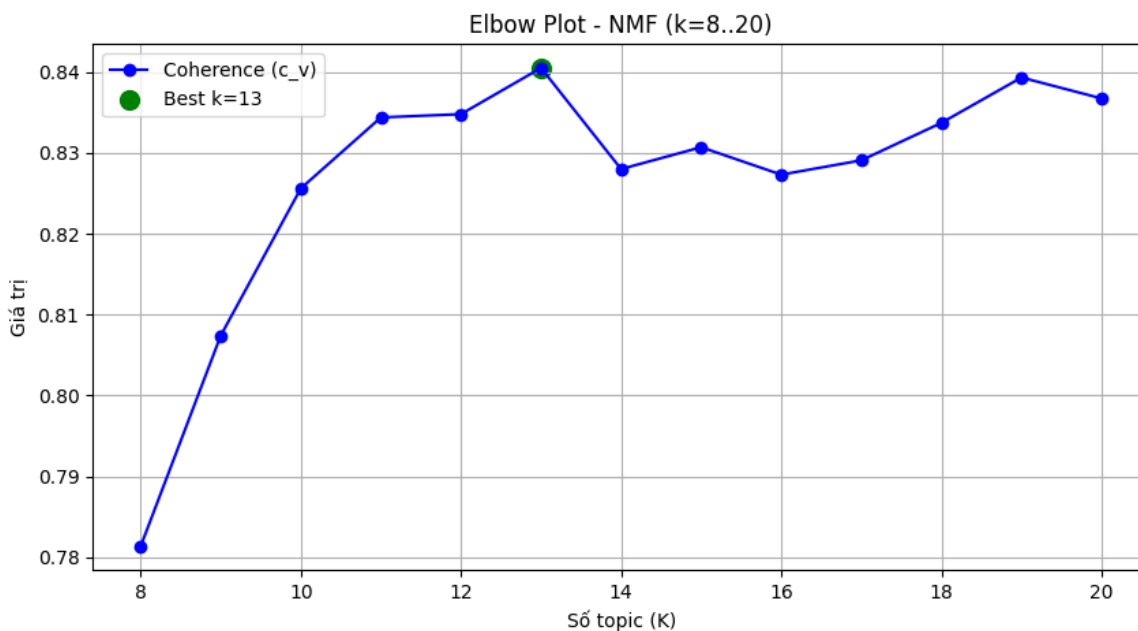
Sau khi thử nghiệm số lượng chủ đề K trong khoảng từ 8 đến 20, kết quả được trình bày trong Bảng III. Chỉ số *Coherence* (c_v) tăng dần từ $K = 8$ đến $K = 13$ và đạt giá trị cao nhất tại $K = 13$ (0.8405). Đồng thời, lỗi tái tạo (*Reconstruction*

Error) giảm dần theo số chủ đề, nhưng mức giảm chậm lại sau ngưỡng $K = 13$. Do đó, $K = 13$ được chọn làm số chủ đề tối ưu cho NMF.

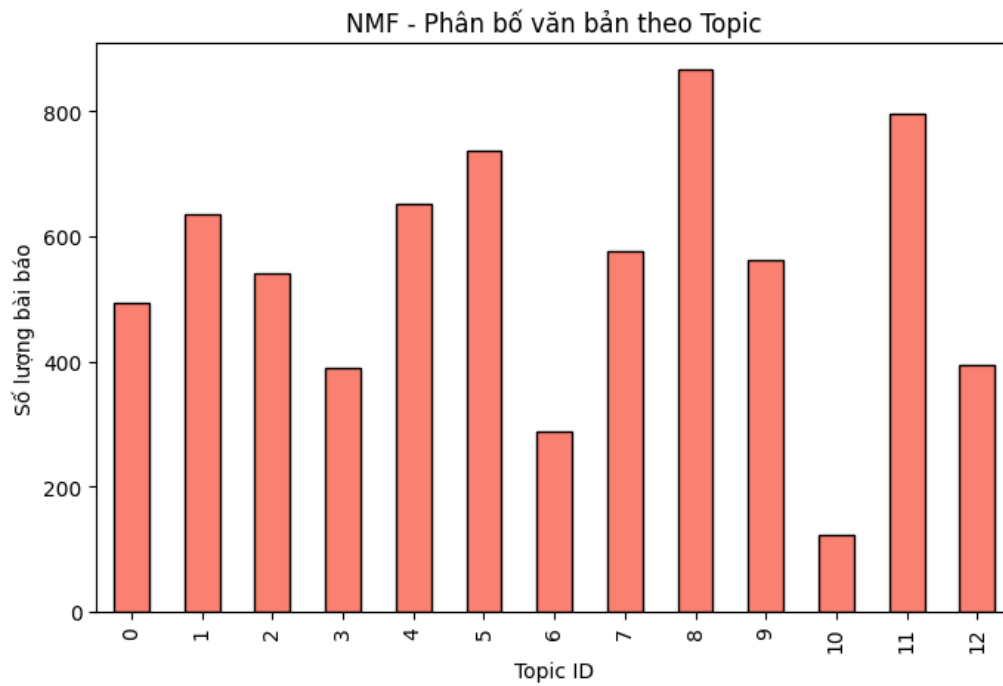
Bảng III
KẾT QUẢ THỬ NGHIỆM NMF VỚI NHIỀU SỐ CHỦ ĐỀ ($K=8..20$)

K	Coherence (c_v)	Reconstruction Error
8	0.7813	81.09
9	0.8074	80.88
10	0.8256	80.71
11	0.8344	80.54
12	0.8348	80.37
13	0.8405	80.24
14	0.8280	80.10
15	0.8307	79.96
16	0.8273	79.87
17	0.8291	79.73
18	0.8337	79.61
19	0.8393	79.50
20	0.8367	79.35

Kết quả này cũng được minh họa trực quan qua Hình 5 với biểu đồ *Elbow Plot*, trong đó điểm cực đại của coherence tại $K = 13$ được đánh dấu. Hình 6 thể hiện phân bố văn bản theo từng chủ đề với $K = 13$, cho thấy dữ liệu được phân bổ khá đồng đều, không bị lệch mạnh vào một topic duy nhất như ở LSA.



Hình 5. Elbow Plot – NMF ($k=8..20$)



Hình 6. Phân bố văn bản theo từng topic (NMF, $K = 13$)

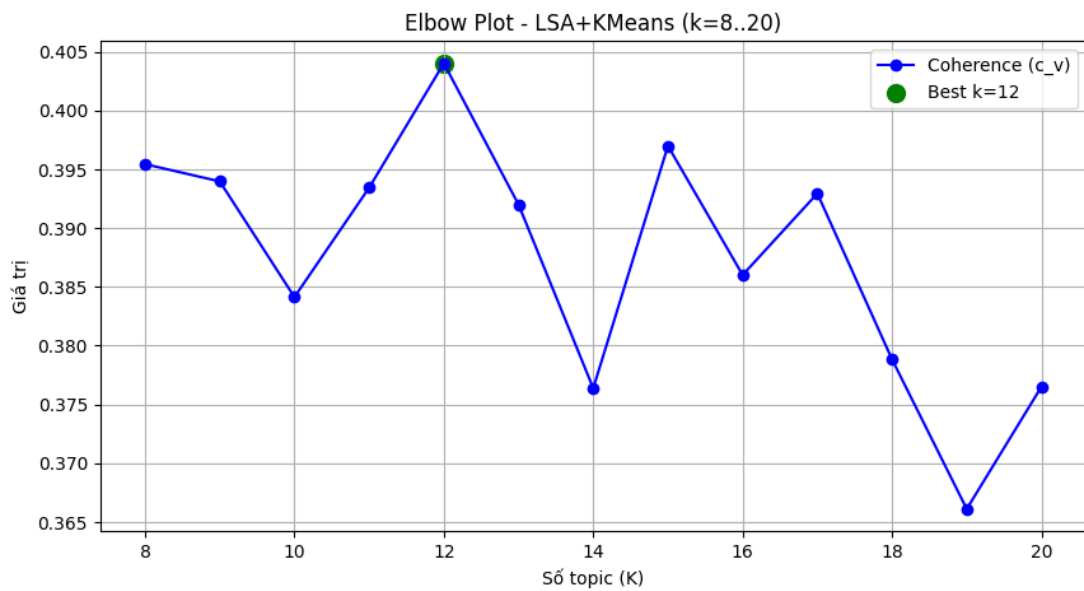
Các chủ đề chính trích xuất từ NMF ($K = 13$):

- **Topic 0 – Kinh tế và doanh nghiệp:** doanh_nghiệp, phát_triển, công_nghệ, bền_vững, thị_trường, tài_chính...
- **Topic 1 – Đời sống gia đình:** chồng, vợ, mẹ, hôn_nhân, yêu, bố_mẹ, sống...
- **Topic 2 – Ô tô và xe điện:** xe, mẫu, ô_tô, vinfast, toyota, xe_điện, phiên_bản...
- **Topic 3 – Thể thao (Tennis):** đấu, trận, giải, vợt, vô_địch, alcaraz, djokovic...
- **Topic 4 – Giáo dục và đào tạo:** trường, học_sinh, đại_học, sinh_viên, thi, phụ_huynh...
- **Topic 5 – An ninh và pháp luật:** công_an, hành_vì, điều_tra, vi_phạm, án, luật_sư...
- **Topic 6 – Quan hệ quốc tế (Nga–Ukraine):** ukraine, nga, tên_lửa, tổng_thống, tấn_công, quân_đội...
- **Topic 7 – Y tế và bệnh lý:** ung_thư, bệnh_nhân, bác_sĩ, điều_trị, phẫu_thuật, thuốc...
- **Topic 8 – Văn hóa và giải trí:** phim, diễn_viên, khán_giả, nghệ_sĩ, ca_sĩ, điện_ảnh...
- **Topic 9 – Tài chính và chứng khoán:** đồng, tỷ, cổ_phiếu, usd, vàng, giao_dịch...
- **Topic 10 – Bóng đá Đông Nam Á:** malaysia, fifa, đội_tuyển, afc, bóng_đá...
- **Topic 11 – Thiên tai và môi trường:** bão, mưa, lũ, ngập, sông, xã, khu_vực...
- **Topic 12 – Chính trị và Đảng:** tỉnh, đại_hội, ủy, đảng_bộ, bí_thư, trung_ương...

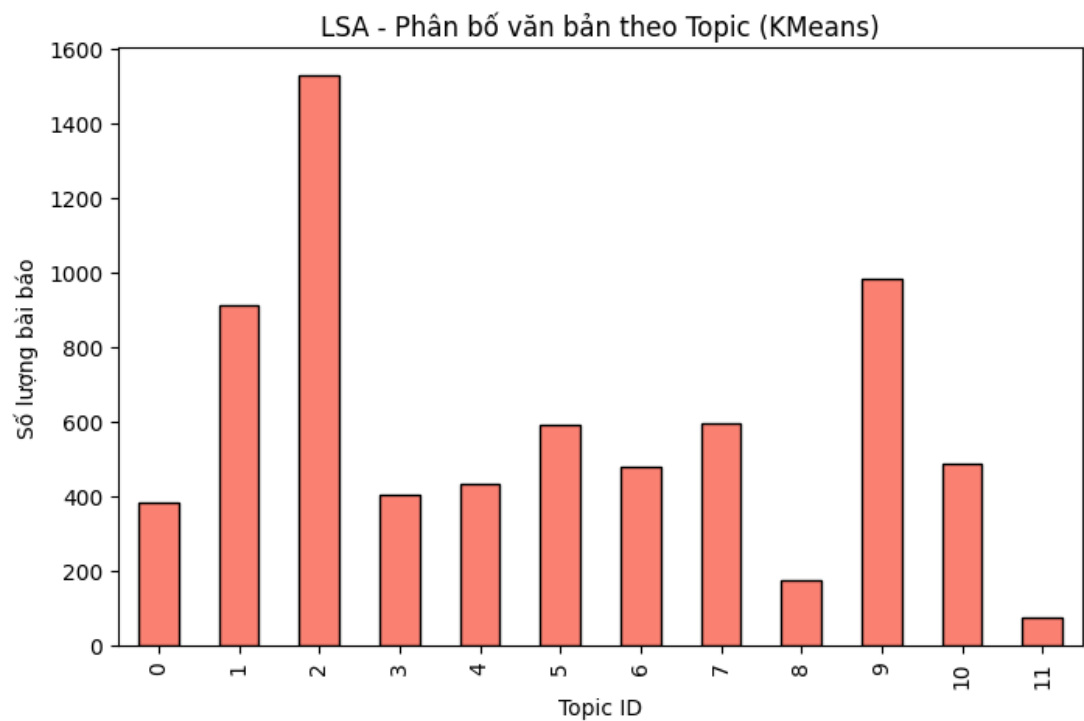
Nhận xét: NMF đạt coherence cao nhất (0.8405) so với LDA và LSA, đồng thời phân bố văn bản giữa các topic khá cân bằng. Chủ đề trích xuất rõ ràng, dễ gắn nhãn và phản ánh đúng các lĩnh vực nội dung chính trong báo điện tử Dân Trí. Điều này cho thấy NMF là một trong những mô hình phù hợp nhất để khai phá chủ đề trong tập dữ liệu nghiên cứu này.

3. LSA

Sau khi áp dụng mô hình **Latent Semantic Analysis (LSA)** kết hợp với thuật toán **KMeans**, kết quả cho thấy số lượng chủ đề tối ưu đạt được là $K = 12$ với chỉ số **coherence** cao nhất $c_v = 0.4040$ và phương sai giải thích (*explained variance*) đạt 0.0681.



Hình 7. Elbow Plot – LSA (k=8..20)



Hình 8. Phân bố văn bản theo topic – LSA ($K = 12$)

Kết quả huấn luyện:

- LSA+KMeans | k=8 → coherence=0.3954 | explained_variance=0.0502
- LSA+KMeans | k=9 → coherence=0.3940 | explained_variance=0.0554
- LSA+KMeans | k=10 → coherence=0.3841 | explained_variance=0.0600
- LSA+KMeans | k=11 → coherence=0.3934 | explained_variance=0.0641
- LSA+KMeans | k=12 → **coherence=0.4040** | explained_variance=0.0681
- LSA+KMeans | k=13 → coherence=0.3920 | explained_variance=0.0717

- LSA+KMeans | k=14 → coherence=0.3764 | explained_variance=0.0750
- LSA+KMeans | k=15 → coherence=0.3970 | explained_variance=0.0783
- LSA+KMeans | k=16 → coherence=0.3860 | explained_variance=0.0814
- LSA+KMeans | k=17 → coherence=0.3929 | explained_variance=0.0843
- LSA+KMeans | k=18 → coherence=0.3788 | explained_variance=0.0872
- LSA+KMeans | k=19 → coherence=0.3661 | explained_variance=0.0899
- LSA+KMeans | k=20 → coherence=0.3765 | explained_variance=0.0927

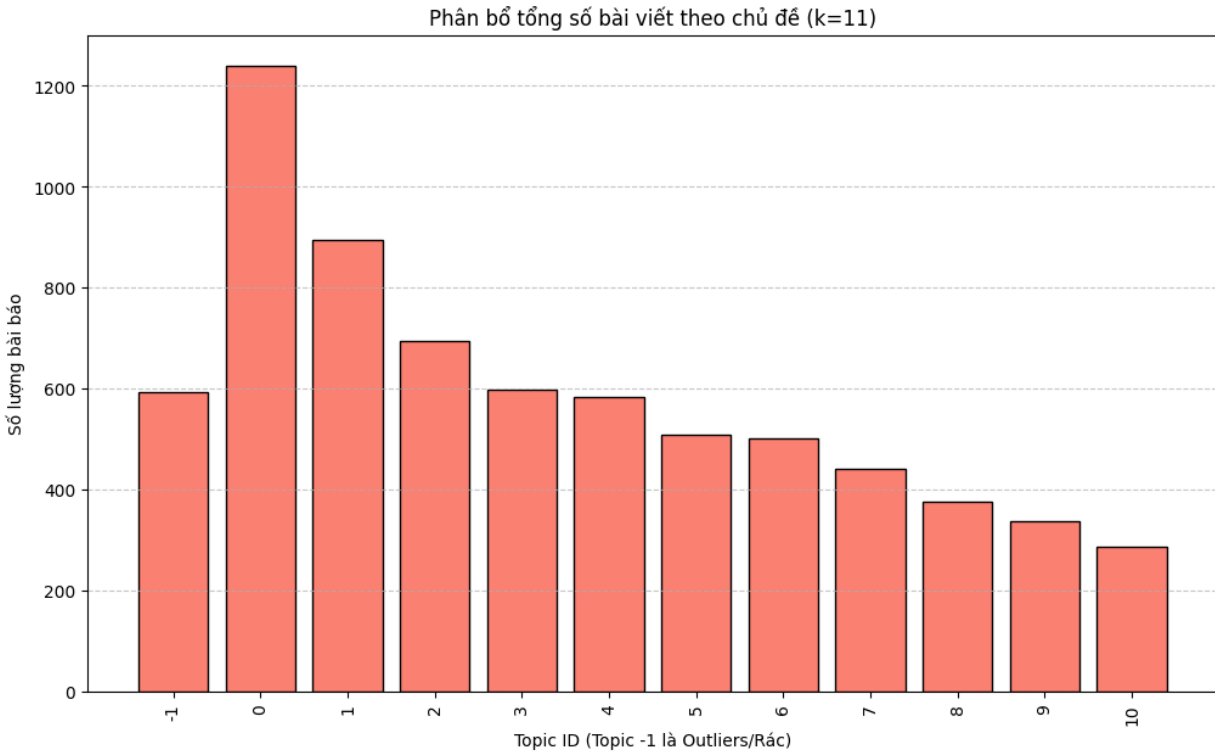
Các chủ đề chính trích xuất từ LSA ($K = 12$):

- **Topic 00:** đại_học, tỉnh, học, tiền, triệu, học_sinh, trường, đồng, việt_nam, xe...
- **Topic 01:** sống, đàn_ông, bố_mẹ, yêu, hôn_nhân, câu_chuyện, chồng, vợ, mẹ, trai...
- **Topic 02:** sinh_viên, công_an, ô_tô, tài_xế, giáo_dục, đại_học, học_sinh, trường, xe...
- **Topic 03:** học_sinh, cầu_thủ, malaysia, hành_vì, giải, đội_tuyển, công_an, xe...
- **Topic 04:** malaysia, đội_tuyển, giáo_dục, thi, học, đại_học, nga, ukraine, xe...
- **Topic 05:** uav, công_an, tên_lửa, fifa, cầu_thủ, malaysia, đội_tuyển, nga, ukraine...
- **Topic 06:** thi_trường, lao_động, công_ty, tỷ_đồng, doanh_nghiệp, trường, học_sinh, ukraine...
- **Topic 07:** ung_thư, bệnh_nhân, điều_trị, bệnh, bác_sĩ, bệnh_viện, y_tế, phẫu_thuật...
- **Topic 08:** phim, diễn_viên, khán_giả, vai, nhân_vật, ca_sĩ, nghệ_thuật, bệnh_nhân, fifa...
- **Topic 09:** đại_hội, sông, người_dân, bão, tỉnh, phát_triển, bị_cáo, đồng, phim...
- **Topic 10:** trận, vdv, giải_đấu, bóng_đá, vô_địch, phim, cầu_thủ, malaysia...
- **Topic 11:** fifa, cầu_thủ, ngập, lao_động, doanh_nghiệp, malaysia, sông, lũ, mưa...

Nhận xét: Mặc dù coherence của LSA ($c_v = 0.4040$) thấp hơn nhiều so với NMF ($c_v = 0.8405$) hay LDA ($c_v = 0.6642$), phương pháp này vẫn cho thấy khả năng gom nhóm văn bản theo ngữ nghĩa ẩn. Tuy nhiên, các topic sinh ra từ LSA có sự chồng chéo lớn, nhiều từ khóa bị pha trộn giữa các lĩnh vực khác nhau (ví dụ: “công_an” xuất hiện cả trong Topic 02, 03, 05). Điều này phản ánh hạn chế của LSA khi áp dụng cho dữ liệu tiếng Việt: không tận dụng tốt được ngữ nghĩa ngữ cảnh như các phương pháp hiện đại hơn.

4. BERTopic

Mô hình **BERTopic** tận dụng sức mạnh của *transformer embeddings* (SBERT) kết hợp với **HDBSCAN** để tự động xác định số lượng chủ đề tối ưu. Kết quả thực nghiệm cho thấy BERTopic trích xuất được tổng cộng $K = 11$ chủ đề chính (trong đó Topic -1 là nhóm outliers/rác). Chỉ số **coherence** đạt $c_v = 0.7324$, cao hơn LDA ($c_v = 0.6642$) và LSA ($c_v = 0.4040$), nhưng thấp hơn NMF ($c_v = 0.8405$). Điều này cho thấy BERTopic khai thác ngữ nghĩa ngữ cảnh tốt nhờ embeddings từ SBERT, đồng thời đảm bảo sự đa dạng chủ đề.



Hình 9. Phân bố tổng số bài viết theo chủ đề – BERTopic ($K = 11$)

Các chủ đề chính trích xuất từ BERTopic:

- **Topic 0 – Văn hóa & Giải trí:** phim, việt_nam, ảnh, khán_giả, diễn_viên, du_lịch, văn_hóa...
- **Topic 1 – An ninh & Pháp luật:** công_an, xe, hành_vì, vụ, bị_cáo, án, tài_xế...
- **Topic 2 – Giáo dục:** trường, đại_học, học_sinh, thi, giáo_dục, giáo_viên, sinh_viên...
- **Topic 3 – Y tế & Sức khỏe:** ung_thư, bệnh, bệnh_nhân, bệnh_viện, điều_trị, bác_sĩ...
- **Topic 4 – Đời sống gia đình:** chồng, vợ, mẹ, gái, hôn_nhân, gia_đình, câu_chuyện...
- **Topic 5 – Thiên tai & Môi trường:** đất, mưa, xã, sông, bão, lũ, ngập, tỉnh...
- **Topic 6 – Thể thao:** đấu, giải, trận, đội_tuyển, thi_đấu, cầu_thủ, malaysia, vdv...
- **Topic 7 – Kinh tế & Doanh nghiệp:** doanh_nghiệp, phát_triển, việt_nam, esg, bền_vững...
- **Topic 8 – Ô tô & Công nghệ:** xe, mẫu, triệu, giá, sạc, vinfast, phiên_bản...
- **Topic 9 – Việc làm & Lao động:** lao_động, công_ty, việc_làm, công_việc, lương, trợ_cấp...
- **Topic 10 – Quan hệ quốc tế & Quân sự:** nga, ukraine, tổng_thống, trump, tên_lửa, uav, moscow...

Nhận xét: BERTopic cho thấy khả năng phân tách chủ đề tốt, gần với trực giác thực tế. Các chủ đề được hình thành khá rõ ràng, ít bị chồng chéo so với LSA. Đặc biệt, nhóm về *đời sống gia đình* (Topic 4) và *xe/ô tô* (Topic 8) được mô hình phát hiện rõ rệt – trong khi ở LDA hoặc NMF, chúng thường bị pha trộn vào các nhóm khác. Tuy nhiên, hạn chế là xuất hiện **Topic -1 (outliers)** với khoảng 600 bài viết không thuộc cụm chủ đề nào, phản ánh tính nhạy cảm của HDBSCAN với nhiễu dữ liệu.

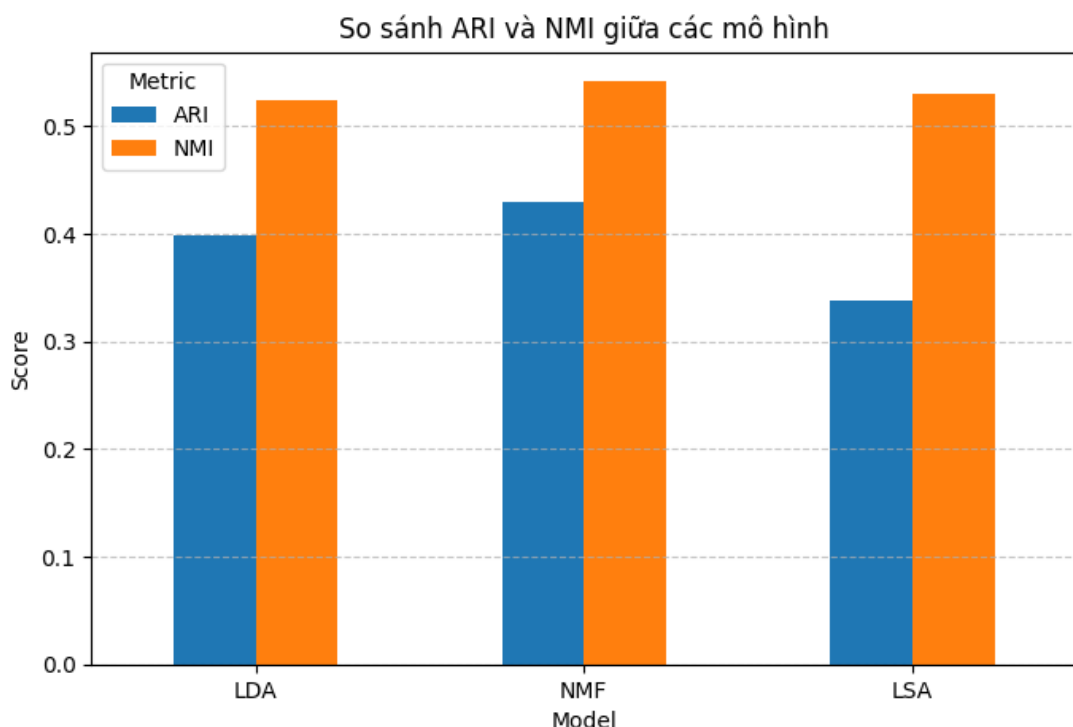
5. ARI và NMI

Để đánh giá chất lượng phân cụm của các mô hình topic modeling, chúng tôi sử dụng hai chỉ số đo lường là Adjusted Rand Index (ARI) và Normalized Mutual Information (NMI). Các chỉ số này đo lường mức độ tương đồng giữa kết quả phân cụm của mô hình và các nhãn thực tế (ground truth). Kết quả được trình bày chi tiết trong Bảng IV và biểu đồ 10.

Model	ARI	NMI
LDA	0.3989	0.5234
NMF	0.4295	0.5417
LSA	0.3373	0.5300

Bảng IV
KẾT QUẢ SO SÁNH CHỈ SỐ ARI VÀ NMI CỦA CÁC MÔ HÌNH.

Dựa trên kết quả từ Bảng IV và biểu đồ 10, ta có thể rút ra các nhận xét sau:



Hình 10. Biểu đồ so sánh chỉ số ARI và NMI giữa các mô hình LDA, NMF và LSA.

- **NMF** là mô hình đạt hiệu suất cao nhất ở cả hai chỉ số, với $ARI = 0.4295$ và $NMI = 0.5417$. Điều này cho thấy khả năng phân cụm của NMF là tốt nhất và gần với thực tế nhất trong ba mô hình được so sánh.
- **LDA** đứng ở vị trí thứ hai, với $ARI = 0.3989$ và $NMI = 0.5234$, cho thấy hiệu suất phân cụm tốt nhưng vẫn kém hơn NMF.
- **LSA** có chỉ số ARI thấp nhất ($ARI = 0.3373$), cho thấy sự tương đồng thấp nhất với các cụm thực tế. Tuy nhiên, chỉ số NMI của LSA ($NMI = 0.5300$) lại cao hơn LDA một chút, cho thấy một góc nhìn khác về chất lượng thông tin tương hỗ. Nhìn chung, NMF cho thấy sự vượt trội rõ rệt trong bài toán phân cụm này, như được minh họa trong biểu đồ so sánh.

VI. ỨNG DỤNG VÀ Ý NGHĨA THỰC TIỄN

Kết quả nghiên cứu về khai phá chủ đề từ dữ liệu báo chí trực tuyến không chỉ mang giá trị học thuật mà còn có nhiều ứng dụng thực tiễn quan trọng:

- **Phân tích báo chí và truyền thông:** Việc trích xuất chủ đề giúp xác định các xu hướng thông tin nổi bật (như chính trị, kinh tế, y tế, giáo dục, văn hoá). Điều này hỗ trợ các cơ quan báo chí định hướng nội dung và theo dõi mối quan tâm của độc giả.
- **Hỗ trợ hoạch định chính sách:** Các nhà quản lý có thể dựa vào kết quả phân tích để nhận diện những vấn đề xã hội được dư luận quan tâm, từ đó đưa ra quyết định chính sách phù hợp.
- **Ứng dụng trong lĩnh vực doanh nghiệp:** Các công ty truyền thông, quảng cáo có thể tận dụng mô hình để phân loại tự động bài viết theo chủ đề, phục vụ cho hệ thống gợi ý (recommendation systems) hoặc nhắm chọn quảng cáo theo ngữ cảnh (contextual advertising).
- **Ứng dụng trong nghiên cứu học thuật:** Các phương pháp so sánh mô hình (LDA, NMF, LSA, BERTopic) cung cấp một khung chuẩn để áp dụng trong nhiều ngữ cảnh khác như phân tích mạng xã hội, bình luận sản phẩm, hoặc dữ liệu khảo sát.

Như vậy, kết quả của đề tài không chỉ dừng lại ở mức độ thử nghiệm, mà còn mở ra nhiều khả năng ứng dụng trong các hệ thống thực tiễn đòi hỏi xử lý khối lượng lớn văn bản tiếng Việt.

VII. HẠN CHẾ VÀ HƯỚNG NGHIÊN CỨU TƯƠNG LAI

Hạn chế

- **Về dữ liệu:** Dữ liệu chỉ thu thập từ một nguồn duy nhất (báo Dân Trí), nên tính khái quát chưa cao. Một số chuyên mục có số lượng bài viết ít, dù đã oversampling, vẫn có nguy cơ gây nhiễu trong kết quả.

- **Về mô hình:**
 - LDA và NMF phụ thuộc nhiều vào tham số số lượng chủ đề K , phải chọn thủ công bằng coherence.
 - LSA+KMeans cho kết quả coherence thấp, ít ý nghĩa thực tiễn.
 - BERTopic vẫn xuất hiện cụm “outliers” (chủ đề rác) và đòi hỏi nhiều tài nguyên GPU.
- **Về đánh giá:** Các chỉ số như Coherence, Reconstruction Error mới phản ánh khía cạnh toán học, chưa kết hợp nhiều đánh giá định tính từ chuyên gia hoặc người dùng thực tế.

Hướng nghiên cứu tương lai

- **Mở rộng dữ liệu:** Thu thập thêm dữ liệu từ nhiều nguồn báo chí khác (VNExpress, Tuổi Trẻ, Thanh Niên...) hoặc từ mạng xã hội để tăng tính đa dạng.
- **Tối ưu mô hình:**
 - Tích hợp **embedding hiện đại** (PhoBERT, vBERT, mBERT) kết hợp với BERTopic để nâng cao chất lượng chủ đề.
 - Áp dụng **học sâu** (neural topic models như ProdLDA, ETM) để giảm phụ thuộc vào số lượng chủ đề đặt trước.
- **Đánh giá thực nghiệm:** Thực hiện khảo sát người dùng hoặc chuyên gia trong lĩnh vực báo chí để đánh giá tính hợp lý, từ đó hiệu chỉnh lại mô hình.
- **Ứng dụng hệ thống thực tế:** Xây dựng demo web/app phân loại và hiển thị chủ đề báo chí theo thời gian thực, tích hợp công cụ tìm kiếm hoặc gợi ý bài viết.

VIII. KẾT LUẬN

Dựa trên toàn bộ thí nghiệm với tập dữ liệu gồm 7.052 bài báo Dân Trí, bốn hướng mô hình hoá chủ đề (**LDA**, **NMF**, **LSA+KMeans**, **BERTopic**) cho thấy sự khác biệt rõ rệt về chất lượng, tính ổn định và khả năng diễn giải. Bảng V tổng hợp các chỉ số then chốt ở cấu hình tối ưu của từng mô hình.

Bảng V
SO SÁNH MÔ HÌNH Ở CẤU HÌNH TỐI ƯU (CHỈ SỐ NỘI VỊ)

Mô hình	K tối ưu	Coherence (c_v)	Chỉ số phụ	Nhận xét ngắn
LDA (Counts)	13	0.6642	—	Chủ đề “sạch”, tách biệt vừa phải, nhạy với tiền xử lý.
NMF (TF-IDF)	13	0.8405	Rec. Err. ↓ (80.24)	Coherence cao nhất; phân bố cân bằng, dễ gán nhãn.
LSA+KMeans (TF-IDF)	12	0.4040	Var. giải thích 0.0681	Chủ đề chồng chéo; một cụm rất lớn, khó diễn giải.
BERTopic (SBERT)	11	0.7324	Có outliers (-1)	Chủ đề bám ngữ cảnh; có cụm rác/ngoại lệ.

1) LDA (ma trận đếm)

Chỉ số: Coherence đạt 0.6642 ở $K=13$. **Kết quả:** Các chủ đề “sạch”, dễ diễn giải: Thời sự/An ninh, Thể thao, Giáo dục, Y tế, Văn hoá... Phân bố khá đều, không có cụm áp đảo. **Ưu/nhược:** Ưu điểm là dễ hiểu và phổ biến, nhưng nhạy với tiền xử lý và coherence thấp hơn NMF/BERTopic.

2) NMF (ma trận TF-IDF)

Chỉ số: Coherence cao nhất trong các mô hình, đạt **0.8405** ở $K=13$; Reconstruction Error giảm dần và ổn định tại $K=13$. **Kết quả:** Chủ đề rõ ràng, sắc nét, ví dụ: Kinh tế/ESG, Đời sống, Ô tô–Xe điện, Thể thao, Giáo dục, Nga–Ukraine, Y tế, Văn hoá. **Ưu/nhược:** Ưu điểm là coherence cao, phân bố cân bằng và dễ gán nhãn. Nhược điểm là không phải mô hình xác suất và cần xác định K thủ công.

3) LSA+KMeans (TF-IDF)

Chỉ số: Coherence thấp (0.4040) ở $K=12$, phương sai giải thích chỉ 0.0681. **Kết quả:** Chủ đề chồng chéo, từ khoá nhiều và khó gán nhãn; một cụm rất lớn và mất cân bằng. **Ưu/nhược:** Ưu điểm là nhanh và đơn giản, nhưng coherence thấp, không phù hợp cho phân tích chủ đề chính xác.

4) BERTopic (SBERT)

Chỉ số: Coherence đạt 0.7324 với 11 chủ đề; xuất hiện một cụm ngoại lệ (-1) gồm ≈ 600 văn bản. **Kết quả:** Chủ đề bám sát ngữ cảnh: Văn hoá/Phim, Giáo dục, Sức khỏe, Gia đình, Thể thao, Kinh tế, Ô tô, Nga–Ukraine... **Ưu/nhược:** Ưu điểm là tận dụng embedding ngữ nghĩa, số chủ đề được xác định tự động. Nhược điểm là cần GPU và có cụm rác cần xử lý.

5) Đánh giá ngoại vi (ARI và NMI)

Ngoài các chỉ số nội vi (intrinsic) như Coherence, chúng tôi bổ sung góc nhìn bằng cách sử dụng các chỉ số ngoại vi (extrinsic) là **Adjusted Rand Index (ARI)** và **Normalized Mutual Information (NMI)** để so sánh kết quả phân cụm của mô hình với các nhãn thực tế (ground truth) của dữ liệu. Phép đo này được áp dụng cho ba mô hình LDA, NMF và LSA.

Bảng VI
SO SÁNH CHỈ SỐ NGOẠI VI ARI VÀ NMI

Mô hình	ARI	NMI
LDA	0.3989	0.5234
NMF	0.4295	0.5417
LSA+KMeans	0.3373	0.5300

Phân tích:

- **NMF** tiếp tục khẳng định vị thế là mô hình hiệu quả nhất, đạt cả $ARI = 0.4295$ và $NMI = 0.5417$ cao nhất. Điều này cho thấy cấu trúc phân cụm của NMF không chỉ gắn kết (coherence cao) mà còn *tương đồng nhất với cấu trúc thực tế* của dữ liệu.
- **LDA** theo sát ở vị trí thứ hai, cho thấy khả năng phân cụm tốt.
- **LSA+KMeans** có chỉ số ARI thấp nhất (0.3373), xác nhận cho nhận định rằng cấu trúc cụm của nó lệch xa so với thực tế, củng cố cho việc nó tạo ra chủ đề chồng chéo và mất cân bằng.

Kết quả đánh giá ngoại vi này **củng cố mạnh mẽ** cho kết luận rằng NMF là lựa chọn tối ưu, theo sau là LDA, cho bộ dữ liệu này.

6) Khuyến nghị lựa chọn

- **NMF (K=13)**: là lựa chọn tối ưu nếu ưu tiên coherence, chủ đề dễ diễn giải, và kết quả phân cụm bám sát thực tế (ARI/NMI cao).
- **BERTopic**: phù hợp khi cần mô hình bám sát ngữ nghĩa và số chủ đề tự động (không yêu cầu so sánh với nhãn ground truth).
- **LDA**: phù hợp cho pipeline BoW truyền thống và yêu cầu mô hình xác suất, cho kết quả phân cụm tốt.
- **LSA+KMeans**: chỉ nên dùng để thăm dò nhanh hoặc làm baseline, do cả coherence và ARI/NMI đều thấp.

Kết luận chung: Trên tập báo Dân Trí đã tiền xử lý, **NMF (K=13)** cho chất lượng cao nhất và phân bố ổn định, được xác thực bởi cả chỉ số nội vi (Coherence) và chỉ số ngoại vi (ARI, NMI). **BERTopic** phù hợp khi cần tính ngữ nghĩa linh hoạt. **LDA** đáng tin cậy cho pipeline BoW truyền thống, trong khi **LSA** chỉ hữu ích ở mức tham khảo ban đầu.

TÀI LIỆU

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
- [4] M. Grootendorst, "Bertopic: Neural topic modeling with contextual embeddings," *arXiv preprint arXiv:2203.05794*, 2022.
- [5] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM*, 2015.
- [6] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.
- [7] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," in *AAAI*, 2002.