PROJECT 2C: PHÂN TÍCH CHỦ ĐỀ DỰA TRÊN MÔ TẢ VIDEO YOUTUBE

1st Hà Thế Anh, 2nd Nguyễn Nhật Nam, 3rd Hoàng Quang Minh
and Le Nhat Tung
HUTECH University, Vietnam
{hatheanh012004, nguyennhatnam01012004, hoangquangminh130804}@gmail.com, and lenhattung@hutech.edu.vn

TÓM TẮT NỘI DUNG

YouTube là nền tảng lưu trữ và chia sẻ video có quy mô dữ liệu khổng lồ, nơi các mô tả video phản ánh xu hướng, chủ đề và hành vi của người dùng trên không gian số. Nghiên cứu này tập trung vào việc phân tích chủ đề trong các mô tả video tiếng Việt bằng các mô hình học máy hiện đại, bao gồm **LDA (Latent Dirichlet Allocation)**, **BERTopic** và **Top2Vec**, đồng thời đề xuất mô hình kết hợp **CombinedTM (LDA + BERTopic)** nhằm cải thiện tính ổn định và khả năng diễn giải chủ đề. Dữ liệu được tiền xử lý thông qua các bước làm sạch ngôn ngữ, loại bỏ ký tự nhiễu, từ dừng, emoji và tách từ bằng Underthesea. Các mô hình được huấn luyện, đánh giá và trực quan hóa dựa trên nhiều chỉ số khác nhau để khám phá cấu trúc ngữ nghĩa tiềm ẩn trong dữ liệu YouTube tiếng Việt. Nghiên cứu góp phần chứng minh tính ứng dụng của các kỹ thuật *Topic Modeling* trong khai phá nội dung văn bản và hiểu biết xu hướng người dùng trên nền tảng video trực tuyến.

TỪ KHÓA

Topic Modeling, Video Description, YouTube, Machine Learning, Latent Dirichlet Allocation (LDA), BERTopic, Top2Vec, Contextualized Topic Model (CTM), Dimensionality Reduction, PCA, UMAP, Clustering, HDBSCAN, Vietnamese Language Data.

I. GIỚI THIỆU

Trong bối cảnh chuyển đổi số và sự bùng nổ của dữ liệu trực tuyến, YouTube đã trở thành một trong những nền tảng chia sẻ nội dung lớn nhất thế giới, nơi hàng triệu video được đăng tải mỗi ngày. Các mô tả video không chỉ cung cấp thông tin về nội dung mà còn phản ánh xu hướng, sở thích và hành vi của người dùng trong không gian mạng. Việc phân tích các mô tả này giúp khai thác tri thức tiềm ẩn, hỗ trợ hệ thống gợi ý, quảng cáo và nghiên cứu hành vi xã hội trên nền tảng số.

Tuy nhiên, dữ liệu mô tả video thường tồn tại dưới dạng văn bản phi cấu trúc, chứa nhiều nhiễu như ký tự đặc biệt, biểu tượng cảm xúc, từ dừng hoặc các yếu tố không mang giá trị ngữ nghĩa. Đặc biệt, với tiếng Việt, sự phức tạp trong cấu trúc ngữ pháp, dấu thanh và cách biểu đạt khiến cho việc xử lý và phân tích ngôn ngữ tự nhiên (NLP) trở nên thách thức hơn. Do đó, việc xây dựng một quy trình tiền xử lý chuyên sâu là bước không thể thiếu nhằm đảm bảo độ chính xác và nhất quán của dữ liệu đầu vào.

Nghiên cứu này tập trung vào bài toán **phân tích chủ đề (Topic Modeling)** trên tập dữ liệu mô tả video YouTube tiếng Việt, với mục tiêu phát hiện các cụm chủ đề tiềm ẩn phản ánh cấu trúc ngữ nghĩa trong nội dung. Ba mô hình chủ đề được áp dụng gồm **LDA (Latent Dirichlet Allocation)**, **BERTopic** và **Top2Vec**, đại diện cho ba hướng tiếp cận khác nhau: Thống kê xác suất, embedding ngữ nghĩa và học chủ đề không giám sát. Bên cạnh đó, nghiên cứu còn đề xuất mô hình lai **CombinedTM (LDA + BERTopic)** nhằm kết hợp ưu điểm của hai mô hình truyền thống và hiện đại, hướng đến tăng cường độ ổn định và khả năng diễn giải của kết quả.

Toàn bộ dữ liệu được thu thập và xử lý theo một quy trình chặt chẽ: Loại bỏ nhiễu, chuẩn hóa ký tự, xóa emoji, hashtag, từ dừng đặc thù của YouTube và tách từ bằng thư viện Underthesea. Sau đó, các mô hình được huấn luyện và đánh giá thông qua nhiều chỉ số khác nhau như **Coherence (C_V, U_Mass, NPMI)**, **Topic Diversity**, **NMI**, **ARI**, **Purity** và **Stability**. Kết hợp cùng các phương pháp trực quan hóa như *WordCloud* và *Bubble Chart*, nghiên cứu mang đến cái nhìn trực quan và toàn diện về cấu trúc chủ đề trong dữ liệu văn bản tiếng Việt trên YouTube.

Công trình này không chỉ góp phần minh chứng hiệu quả của các mô hình phân tích chủ đề hiện đại trong ngôn ngữ tiếng Việt mà còn mở ra hướng tiếp cận mới trong việc khai phá tri thức, phân loại nội dung và nhận diện xu hướng trên các nền tảng video trực tuyến.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong lĩnh vực **phân tích chủ đề (Topic Modeling)**, nhiều công trình nghiên cứu đã được phát triển nhằm khám phá cấu trúc tiềm ẩn của văn bản. Từ các phương pháp thống kê truyền thống đến các mô hình ngữ nghĩa hiện đại, mục tiêu chung là trích xuất được các cụm từ đại diện phản ánh ngữ nghĩa tổng quát của dữ liệu. Phần này trình bày các công trình tiêu biểu có liên quan đến đề tài, bao gồm bốn hướng chính: LDA, BERTopic, Top2Vec và các mô hình kết hợp (CTM).

A. Latent Dirichlet Allocation (LDA)

Mô hình **LDA** do [1] đề xuất là một trong những nền tảng quan trọng nhất của Topic Modeling. LDA giả định rằng mỗi tài liệu là sự pha trộn của nhiều chủ đề, và mỗi chủ đề là một phân bố xác suất trên các từ vựng. Phương pháp này sử dụng các kỹ thuật xác suất Bayes để suy luận phân bố tiềm ẩn của các chủ đề trong tập dữ liệu văn bản. Các khảo sát như của [2] đã chỉ ra rằng LDA có độ ổn định cao và khả năng diễn giải tốt, nhưng lại phụ thuộc nhiều vào tiền xử lý dữ liệu, đặc biệt trong các ngôn ngữ có cấu trúc phức tạp như tiếng Việt.

B. BERTopic - Mô hình chủ đề dựa trên ngữ nghĩa

BERTopic, được phát triển bởi [3], là mô hình hiện đại kết hợp *sentence embedding*, gom cụm bằng HDBSCAN và trích xuất từ khóa bằng c-TF-IDF. Khác với LDA dựa trên tần suất từ, BERTopic khai thác thông tin ngữ nghĩa từ các mô hình ngôn ngữ Transformer (như BERT, RoBERTa), giúp nhóm các văn bản có ngữ cảnh tương tự lại với nhau. Theo [4], BERTopic đạt hiệu suất vượt trội trong việc phân tích dữ liệu mạng xã hội và nội dung đa ngôn ngữ nhờ khả năng học biểu diễn ngữ nghĩa sâu.

C. Top2Vec - Học chủ đề trên không gian vector

Top2Vec, do [5] giới thiệu, là một hướng tiếp cận mới trong học biểu diễn chủ đề. Phương pháp này học trực tiếp embedding của từ, tài liệu và chủ đề trong cùng một không gian vector, qua đó tự động phát hiện và gom nhóm các tài liệu mà không cần xác định trước số lượng chủ đề. Top2Vec khắc phục một số hạn chế của LDA khi làm việc với văn bản ngắn, tuy nhiên mô hình có thể gặp khó khăn với ngôn ngữ ít tài nguyên hoặc dữ liệu nhiễu như tiếng Việt.

D. Combined Topic Models (CTM)

Để kết hợp sức mạnh của các phương pháp thống kê và embedding ngữ nghĩa, nhiều công trình gần đây đã đề xuất **Combined Topic Models (CTM)**. Theo [6], việc tích hợp biểu diễn xác suất của LDA với embedding từ BERT giúp cải thiện đáng kể tính mạch lạc (coherence) và khả năng diễn giải của các chủ đề. Ngoài ra, nghiên cứu của [7] đã phát triển thư viện `contextualized-topic-models` cho phép huấn luyện các mô hình CTM trên dữ liệu đa ngôn ngữ, thể hiện hiệu quả cao trong các bài toán phân tích ngữ nghĩa phức tạp. Những ý tưởng này là nền tảng để xây dựng mô hình **CombinedTM (LDA + BERTopic)** trong nghiên cứu hiện tại, nhằm tận dụng đồng thời khả năng tổng quát hóa của LDA và sức mạnh biểu diễn ngữ cảnh của BERTopic.

E. Các khảo sát và ứng dụng gần đây

Theo tổng quan của [8], việc áp dụng các mô hình phân tích chủ đề trong dữ liệu mạng xã hội và mô tả ngắn (như video YouTube) đang trở thành xu hướng nghiên cứu quan trọng. Các khảo sát này nhấn mạnh vai trò của việc kết hợp mô hình truyền thống và hiện đại để đạt được sự cân bằng giữa độ ổn định và độ chính xác ngữ nghĩa. Bên cạnh đó, [4] cũng chỉ ra rằng các mô hình embedding như BERTopic và Top2Vec thường đạt hiệu quả cao hơn LDA khi làm việc với dữ liệu ngắn và đa chủ đề, nhưng lại cần nhiều tài nguyên tính toán hơn.

F. Tổng kết

Tổng hợp các nghiên cứu trên cho thấy sự tiến hóa của các mô hình phân tích chủ đề theo hướng kết hợp giữa *thống kê xác suất* và *ngữ nghĩa học sâu*. LDA vẫn là nền tảng vững chắc để mô hình hóa xác suất chủ đề, trong khi BERTopic và Top2Vec mở ra khả năng hiểu ngữ nghĩa sâu hơn của văn bản. Các mô hình lai như CombinedTM (LDA + BERTopic) chứng minh tiềm năng lớn trong việc nâng cao độ ổn định, tính mạch lạc và khả năng ứng dụng vào các bài toán khai phá nội dung tiếng Việt.

III. PHƯƠNG PHÁP NGHIÊN CỨU

Phần này mô tả quy trình và các kỹ thuật phân tích chủ đề được áp dụng trong đề tài *"Phân tích chủ đề trên dữ liệu mô tả video YouTube"*. Quy trình được thiết kế thành pipeline tự động, bao gồm bốn khối chính: (1) tiền xử lý văn bản, (2) mô hình hóa chủ đề, (3) đánh giá mô hình và (4) trực quan hóa kết quả.

A. Phương pháp tiền xử lý dữ liệu

Dữ liệu đầu vào được lấy từ tập `Project2C_youtube_data_cleaned.csv`, trong đó mỗi bản ghi chứa trường `clean_description` chuỗi mô tả video đã được làm sạch thông qua hàm `clean_text()` được xây dựng thủ công. Các bước xử lý gồm:

- Chuyển toàn bộ văn bản về dạng chữ thường
- Loại bỏ biểu tượng cảm xúc, URL, thẻ hashtag, ký tự đặc biệt và từ khóa quảng bá đặc trưng của YouTube (*like*, *subscribe*, *share*, ...)
- Chuẩn hóa dấu và ký tự tiếng Việt, chỉ giữ lại chữ cái và khoảng trắng;
- Tách từ bằng thư viện `Underthesea`
- Loại bỏ từ dừng tiếng Việt được mở rộng thủ công
- Lọc bỏ các mô tả ngắn dưới 5 token để đảm bảo đủ ngữ cảnh.

Kết quả thu được là danh sách token hóa (`texts`) và từ điển ánh xạ id–từ (`dictionary`) phục vụ cho quá trình huấn luyện LDA và các mô hình khác.

B. Phương pháp mô hình hóa chủ đề

Ba mô hình chủ đề được triển khai gồm **LDA**, **BERTopic** và **Top2Vec**, bên cạnh đó nhóm nghiên cứu đề xuất mô hình lai **CombinedTM (LDA + BERTopic)** nhằm tận dụng ưu thế của cả hai hướng tiếp cận: xác suất và ngữ nghĩa.

1) Mô hình LDA (Latent Dirichlet Allocation): LDA giả định rằng mỗi tài liệu d là sự pha trộn của các chủ đề tiềm ẩn z, và mỗi chủ đề là một phân bố xác suất trên không gian từ vựng w. Quy trình sinh của mô hình được mô tả như sau:

$$P(w|d) = \sum_{z=1}^{K} P(w|z)P(z|d)$$

trong đó K là số chủ đề được xác định thông qua quá trình thử nghiệm trong khoảng 5 – 40. Việc huấn luyện được thực hiện bằng thư viện `gensim.models.LdaModel` với tham số `passes`=20, `iterations`=100 và `random_state`=42. Hai chỉ số được dùng để chọn mô hình tối ưu là **Coherence** (C_V) và **Perplexity** [9].

2) Mô hình BERTopic: Mô hình **BERTopic** [3] sử dụng embedding ngữ nghĩa từ mô hình Transformer đa ngôn ngữ để biểu diễn các mô tả video trong không gian vector. Quy trình gồm ba giai đoạn: (1) mã hóa câu bằng mô hình Transformer, (2) giảm chiều bằng `UMAP` [10], (3) gom cụm bằng `HDBSCAN` [11]. Các cụm chủ đề được đặt tên bằng phương pháp **class-based TF-IDF (c-TF-IDF)** để trích xuất các từ khóa đại diện cho từng nhóm chủ đề. BERTopic được triển khai thông qua thư viện chính thức tại maartengr.github.io/BERTopic.

3) Mô hình Top2Vec: Mô hình **Top2Vec** [12] huấn luyện embedding của tài liệu, từ và chủ đề trong cùng một không gian vector. Khác với LDA, Top2Vec không yêu cầu xác định trước số lượng chủ đề mà tự động phát hiện các cụm dựa trên mật độ điểm embedding. Quy trình gồm ba bước: (1) sinh embedding bằng mô hình học sâu, (2) giảm chiều bằng `UMAP`, (3) phát hiện cụm chủ đề bằng `HDBSCAN`. Cấu trúc học này giúp phát hiện các cụm ngữ nghĩa tổng thể của dữ liệu và được triển khai bằng thư viện mã nguồn mở tại github.com/ddangelov/Top2Vec.

4) Mô hình CombinedTM (LDA + BERTopic): Để kết hợp khả năng diễn giải cao của LDA [13] và khả năng học ngữ nghĩa sâu của BERTopic [3], một mô hình lai **CombinedTM** được đề xuất. Mỗi tài liệu được biểu diễn dưới dạng vector kết hợp giữa phân bố xác suất chủ đề từ LDA (θ_{LDA}) và embedding giảm chiều từ BERTopic sau khi PCA [14] (PCA(E_{BERTopic})), theo công thức:

$$\text{CombinedEmb} = [\theta_{\text{LDA}} \;||\; \text{PCA}(E_{\text{BERTopic}})]$$

Trong đó, ký hiệu || thể hiện phép nối vector. Các vector được chuẩn hóa bằng `StandardScaler` [15], giảm chiều bằng `UMAP` [10], và gom cụm bằng `HDBSCAN` [11]. Mô hình này tận dụng sức mạnh thống kê của LDA trong việc diễn giải và khả năng học ngữ nghĩa sâu của BERTopic để tạo ra các cụm chủ đề ổn định và giàu ý nghĩa hơn.

IV. Xây dựng và chuẩn bị dữ liệu

A. Nguồn dữ liệu và cấu trúc ban đầu

Tập dữ liệu gốc được thu thập trực tiếp từ nền tảng YouTube thông qua API, bao gồm các thông tin cơ bản của video được công bố tại khu vực Việt Nam. Dữ liệu được lưu trữ dưới định dạng CSV với các trường chính như:

- `video_id`: mã định danh duy nhất của video trên YouTube
- `title`: tiêu đề của video;
- `description`: phần mô tả nội dung do người đăng tải cung cấp
- `published_at`: thời điểm xuất bản video (định dạng ISO 8601)
- `channel`: tên kênh đăng video;

- `category_id`: mã danh mục chủ đề của video
- `tags`: danh sách từ khóa gắn liền với video;
- `view_count`, `like_count`, `comment_count`: các chỉ số tương tác (lượt xem, thích, bình luận).

Các video được lấy mẫu từ nhiều danh mục khác nhau như *Âm nhạc*, *Giải trí*, *Thể thao*, *Khoa học & Công nghệ*, nhằm đảm bảo tính đa dạng nội dung. Danh mục chính thức (`category_name`) được ánh xạ từ mã `category_id` thông qua **YouTube Data API v3** [16], với endpoint:

`https://www.googleapis.com/youtube/v3/videoCategories?part=snippet&regionCode=VN`

Kết quả trả về bao gồm cặp giá trị `id` và `title`, giúp gán tên danh mục tiếng Anh và được dịch sang tiếng Việt trong nghiên cứu.

B. Tiền xử lý dữ liệu mô tả (Description)

Cột `description` được xem là trường dữ liệu quan trọng nhất, đóng vai trò đầu vào cho toàn bộ pipeline phân tích chủ đề. Để đảm bảo chất lượng và loại bỏ nhiễu trong ngôn ngữ tự nhiên, một hàm `clean_text()` được xây dựng nhằm thực hiện các bước xử lý sau:

1) **Chuyển chữ thường và loại bỏ khoảng trắng dư thừa.**
2) **Loại bỏ biểu tượng cảm xúc và ký tự đặc biệt:** Sử dụng thư viện `emoji` [17] để xoá toàn bộ biểu tượng ngoài ngôn ngữ.
3) **Xóa URL, email, hashtag, số điện thoại:** Áp dụng biểu thức chính quy (`regex`) để làm sạch chuỗi.
4) **Tách từ tiếng Việt:** Áp dụng công cụ `Underthesea` [18] để tách từ theo quy tắc ngữ pháp tiếng Việt.
5) **Loại bỏ từ dừng và token ngắn:** Danh sách stopwords mở rộng bao gồm các từ thông dụng và các cụm phổ biến trong video như "đăng ký", "like", "share", "cảm ơn", "video", "kênh".
6) **Xử lý lặp và lọc độ dài:** Loại bỏ token trùng liền kề (ví dụ "haha haha") và chỉ giữ mô tả có ít nhất 5 token.

Kết quả của quá trình này là cột `clean_description`, được thêm vào DataFrame và lưu ra tệp. Tổng cộng có 7653 mô tả hợp lệ sau khi loại bỏ các hàng rỗng hoặc mô tả ngắn.

C. Chuẩn bị dữ liệu cho huấn luyện

Tập dữ liệu sau khi làm sạch được chuẩn bị theo hai cấu trúc chính:

- **Token list (`texts`):** Mỗi phần tử là một danh sách các từ (token) trong mô tả video, phục vụ cho các mô hình dựa trên tần suất như LDA.
- **Corpus và Dictionary:** Sử dụng `gensim.corpora.Dictionary` để ánh xạ từ–ID, và `doc2bow()` để tạo vector BoW (Bag-of-Words). Cấu trúc này là đầu vào của mô hình LDA, trong khi BERTopic và Top2Vec sử dụng dạng chuỗi văn bản gốc.

Tất cả văn bản được mã hóa UTF-8 để đảm bảo tương thích với các thư viện xử lý tiếng Việt. Dữ liệu cuối cùng gồm hai tệp chính:

- `Project2C_youtube_data_cleaned.csv`: chứa mô tả đã tiền xử lý.
- `lda_models/`: thư mục lưu mô hình LDA được huấn luyện cho các giá trị chủ đề khác nhau.

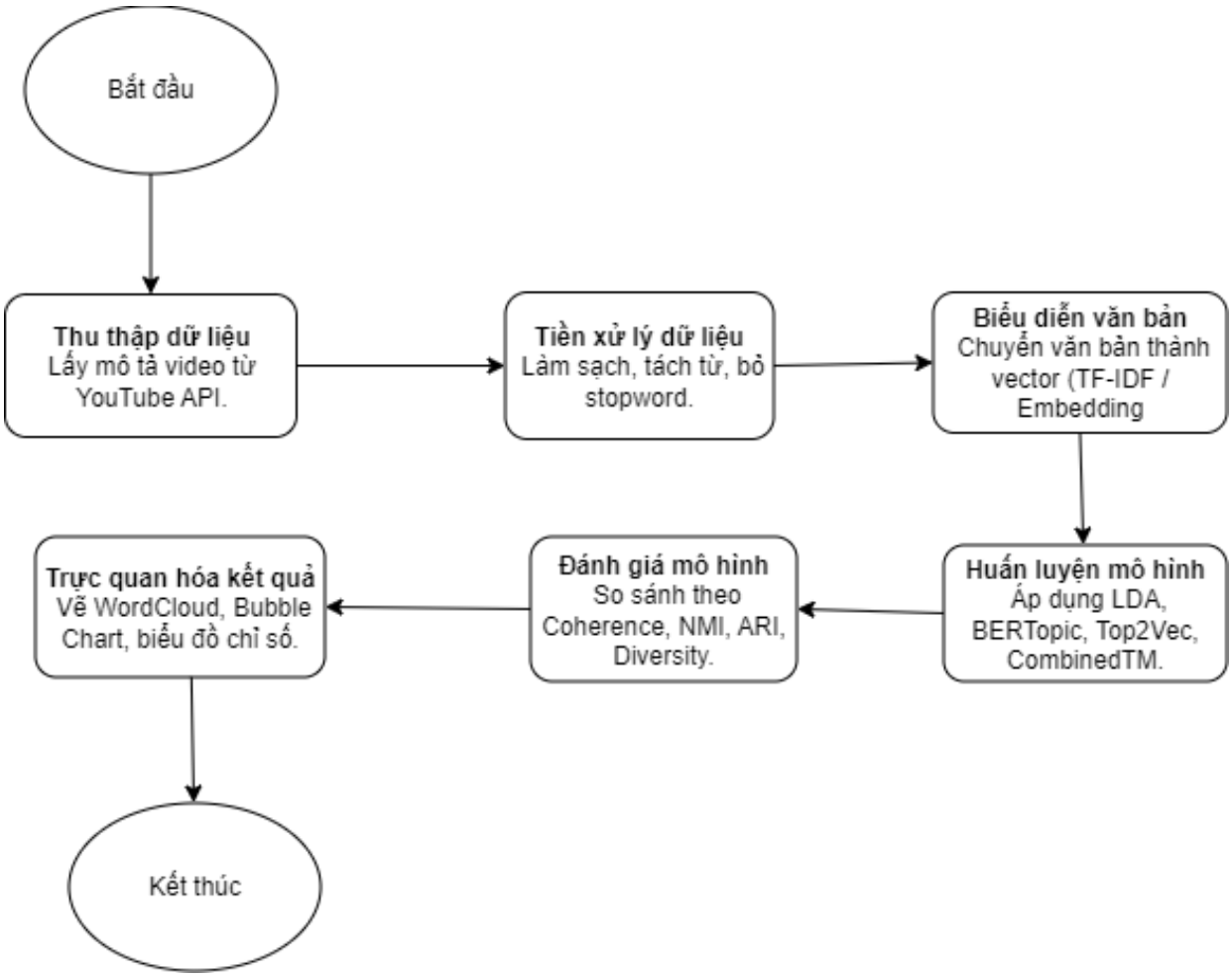
V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Thiết lập thực nghiệm

Các mô hình được triển khai và huấn luyện trực tiếp trên tập dữ liệu mô tả video YouTube sau khi tiền xử lý. Toàn bộ quá trình được thực hiện trong môi trường **Python 3.10** sử dụng các thư viện `gensim`, `bertopic`, `top2vec`, `hdbscan`, `umap-learn`, và `scikit-learn`. Tập dữ liệu gồm **7.653 mô tả video hợp lệ**, sau khi loại bỏ các trường thiếu, ký tự đặc biệt, emoji, đường dẫn URL, và các mô tả ngắn hơn 5 token.

Quy trình thực nghiệm được mô tả trong Hình 1, bao gồm bốn giai đoạn chính:

1) **Tiền xử lý dữ liệu:** làm sạch văn bản, chuẩn hóa, tách từ, loại bỏ stopwords tiếng Việt tự định nghĩa.
2) **Huấn luyện mô hình chủ đề:** lần lượt huấn luyện ba mô hình LDA, BERTopic và Top2Vec trên cùng tập dữ liệu đã làm sạch.
3) **Đánh giá định lượng:** tính toán các chỉ số *Coherence, Perplexity, NMI, ARI, Purity, Topic Diversity*.
4) **Trực quan hóa và phân tích:** biểu diễn kết quả bằng biểu đồ Coherence, Bubble Chart, WordCloud và Bar Plot.



Hình 1. Quy trình thực nghiệm phát hiện và đánh giá chủ đề trên dữ liệu YouTube.

Không áp dụng chia tập huấn luyện và kiểm thử, do mục tiêu của nghiên cứu là khám phá và so sánh chất lượng trích xuất chủ đề giữa các mô hình, thay vì huấn luyện mô hình dự đoán. Mỗi mô hình được huấn luyện một lần duy nhất với cấu hình cố định, đảm bảo tính nhất quán khi so sánh kết quả.

B. Cấu hình tham số mô hình

a) LDA - Latent Dirichlet Allocation.: LDA được huấn luyện trên tập token hoá sử dụng `gensim.LdaModel`. Để xác định số chủ đề tối ưu, mô hình được thử nghiệm với K trong khoảng $[5, 10, 15, 20, 25, 30, 35, 40]$. Mỗi lần huấn luyện chạy 20 vòng (`passes=20`) với 100 vòng lặp nội bộ (`iterations=100`). Chỉ số **Coherence** và **Perplexity** được sử dụng để lựa chọn cấu hình tốt nhất. Kết quả cho thấy số chủ đề tối ưu là **5**, tương ứng với $C_v = 0.549$ và $Perplexity = -7.74$.

b) BERTopic.: Mô hình BERTopic được cấu hình ở chế độ đa ngôn ngữ, cho phép xử lý các mô tả video tiếng Việt một cách hiệu quả. Mô hình sử dụng kiến trúc Transformer để mã hoá câu, sau đó giảm chiều bằng **UMAP** và gom cụm bằng **HDBSCAN**. Từ khóa đại diện cho từng chủ đề được trích xuất bằng phương pháp **class-based TF-IDF (c-TF-IDF)**, giúp tăng khả năng diễn giải và tách biệt ngữ nghĩa giữa các nhóm chủ đề.

c) Top2Vec.: Top2Vec được huấn luyện với tham số `speed='learn'` và `workers=8`. Mô hình kết hợp giữa embedding của tài liệu và cụm chủ đề trong cùng không gian vector, cho phép phát hiện chủ đề dựa trên sự tương đồng cosine của các đoạn mô tả. Các topic được tự động gán cho từng tài liệu thông qua phương thức `get_documents_topics()`.

d) CombinedTM (LDA + BERTopic).: Mô hình kết hợp được thiết kế nhằm tận dụng ưu điểm của LDA (khả năng mô hình hóa xác suất chủ đề) và BERTopic (khả năng học ngữ nghĩa sâu). Cụ thể, embedding của mỗi mô tả được xây dựng như sau:

$$\mathbf{h}_{\text{Combined}} = [\boldsymbol{\theta}_{LDA} \,\|\, \text{PCA}(E_{BERTopic})]$$

trong đó $\boldsymbol{\theta}_{LDA}$ là phân bố xác suất chủ đề của tài liệu trong không gian LDA, và $E_{BERTopic}$ là vector ngữ nghĩa rút ra từ mô hình BERTopic (sau khi giảm chiều bằng PCA). Các vector hợp nhất này được chuẩn hóa bằng `StandardScaler`, sau

đó giảm chiều bằng UMAP và gom cụm bằng HDBSCAN. Mô hình này hướng tới việc tăng tính ổn định và độ chính xác phân cụm chủ đề.

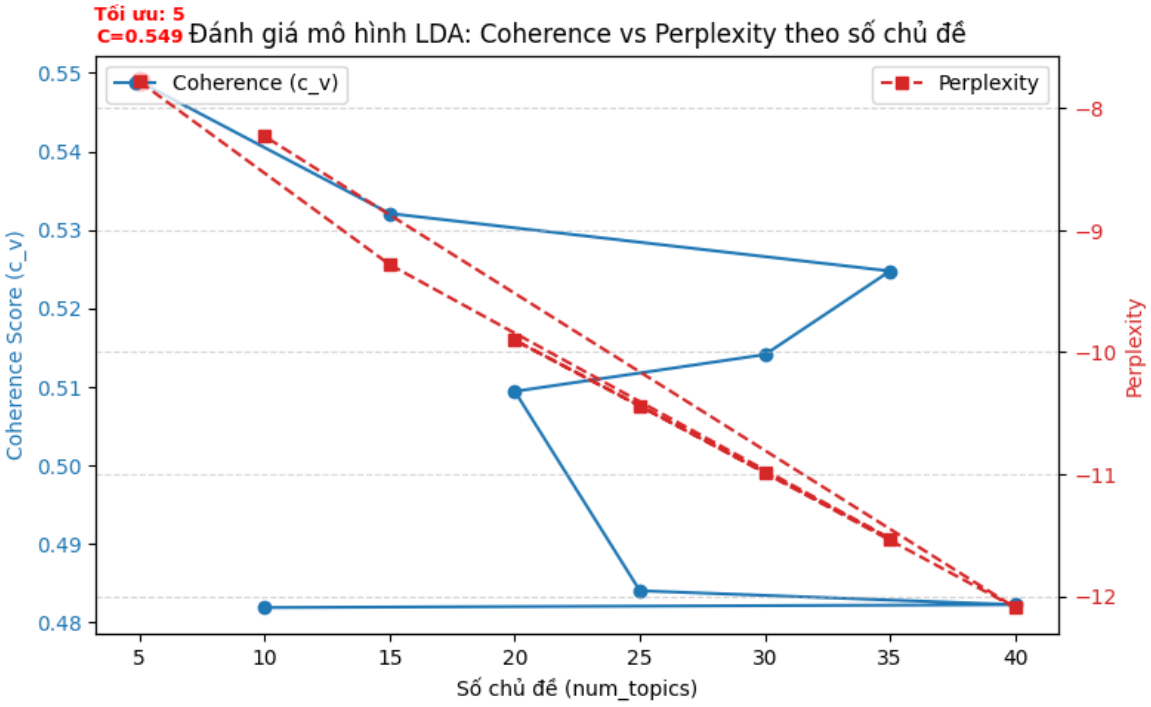
Sau khi thiết lập, các mô hình được đánh giá dựa trên các tiêu chí: **Coherence (C_v, U_Mass, NPMI)**, **NMI**, **ARI**, **Purity**, và **Topic Diversity**. Các kết quả chi tiết và biểu đồ so sánh được trình bày ở các phần tiếp theo.

C. Kết quả và phân tích

Phần này trình bày kết quả thực nghiệm và đánh giá hiệu năng của các mô hình phân tích chủ đề được áp dụng trong nghiên cứu, bao gồm: **LDA**, **BERTopic**, **Top2Vec**, và **CombinedTM**. Các mô hình được huấn luyện trên tập dữ liệu mô tả video YouTube đã được tiền xử lý, với mục tiêu xác định phương pháp nào mang lại hiệu quả cao nhất về mặt ngữ nghĩa và độ phân tách chủ đề.

1) Kết quả mô hình LDA: Mô hình **Latent Dirichlet Allocation (LDA)** được huấn luyện với số lượng chủ đề thay đổi từ 5 đến 40 để xác định cấu hình tối ưu. Hai chỉ số chính được sử dụng trong quá trình đánh giá là **Coherence** và **Perplexity**.

Kết quả cho thấy khi số chủ đề đạt khoảng 5–10, giá trị *Coherence* đạt mức cao nhất trong khi *Perplexity* giảm dần và ổn định (Hình 2). Điều này chứng tỏ mô hình LDA có khả năng phát hiện được cấu trúc chủ đề hợp lý, tránh tình trạng phân mảnh khi số chủ đề quá lớn.

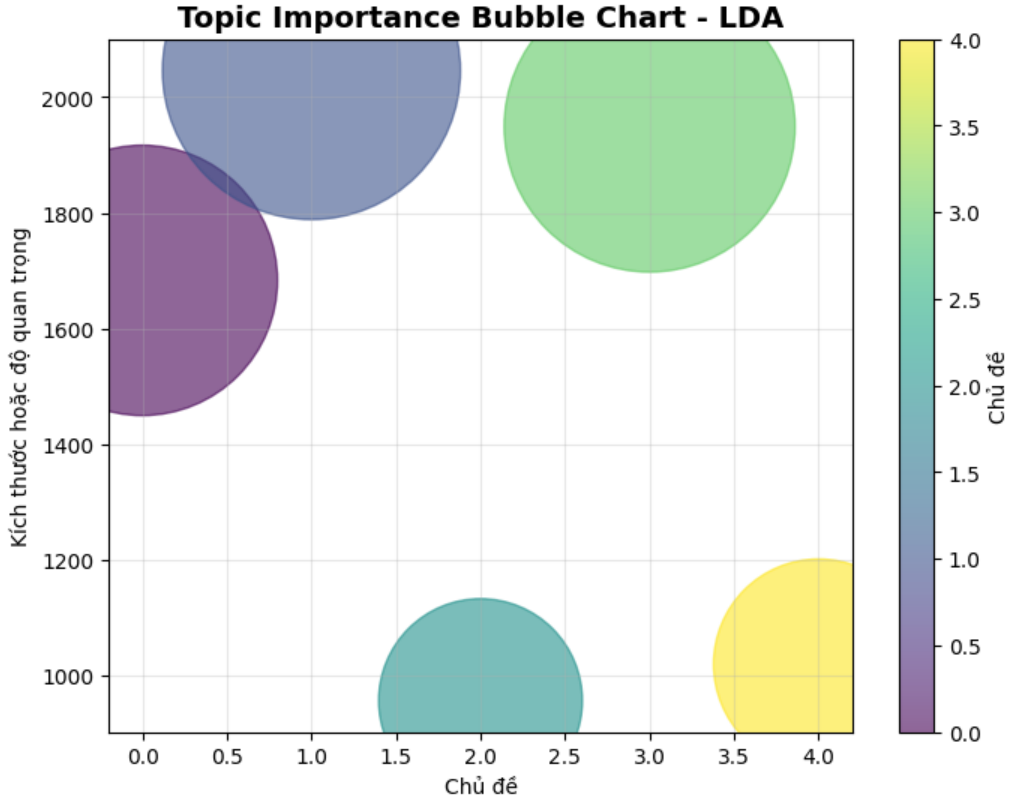


Hình 2. Mối quan hệ giữa số lượng chủ đề và giá trị Coherence/Perplexity của mô hình LDA.

Phân tích nội dung các chủ đề sinh ra cho thấy LDA hình thành được các nhóm từ khóa có liên quan ngữ nghĩa như *"âm nhạc, video, khán giả"*, *"thể thao, bóng đá, đội tuyển"*, hoặc *"du lịch, trải nghiệm, địa điểm"*. Điều này thể hiện khả năng phân cụm theo ngữ cảnh khá tốt của LDA trong dữ liệu YouTube. Tuy nhiên, do đặc trưng dựa trên phân phối từ xác suất, một số chủ đề vẫn xuất hiện hiện tượng *chồng chéo nội dung*, đặc biệt ở các lĩnh vực có ngữ nghĩa gần nhau như *"giải trí"* và *"vlog cá nhân"*.

Hình 3. Biểu đồ WordCloud thể hiện các từ khóa nổi bật trong chủ đề sinh ra bởi LDA.

Nhìn chung, LDA mang lại độ ổn định và khả năng diễn giải trực quan tương đối tốt. Mô hình phù hợp cho các bài toán khám phá chủ đề tổng quát, nhưng hạn chế khi cần biểu diễn các mối quan hệ ngữ nghĩa sâu giữa các từ, điều mà các mô hình embedding hiện đại (như BERTopic) xử lý hiệu quả hơn.



Hình 4. Biểu đồ Bubble Chart thể hiện tầm quan trọng của các chủ đề trong mô hình LDA.

Hình 4 thể hiện kích thước và tầm quan trọng tương đối của từng chủ đề trong không gian hai chiều. Các chủ đề được biểu diễn dưới dạng các bong bóng có kích thước khác nhau, trong đó kích thước phản ánh độ phổ biến của chủ đề, và vị trí thể hiện mối quan hệ ngữ nghĩa tương đối giữa các cụm.

Quan sát cho thấy năm chủ đề chính có kích thước tương đối đồng đều, thể hiện khả năng phân bổ xác suất hợp lý của LDA. Không có cụm nào chiếm ưu thế vượt trội, cho thấy mô hình giữ được sự cân bằng giữa các nhóm nội dung. Tuy nhiên,

do giới hạn trong biểu diễn ngữ nghĩa, các cụm vẫn có mức độ chồng lấn nhất định, đặc biệt ở các chủ đề liên quan đến *"văn hóa", "giải trí"* và *"truyền thông"*, vốn có xu hướng sử dụng từ ngữ tương đồng.

Về mặt định lượng, LDA đạt giá trị **Coherence** cao nhất ở mức $C\_v = 0.549$ với $num\_topics = 5$, và **Perplexity** đạt giá trị thấp nhất ở khoảng -7.7. Các chỉ số này cho thấy mô hình đã tìm được số chủ đề tối ưu, đảm bảo cân bằng giữa tính khái quát và khả năng giải thích. Dù không đạt hiệu quả cao như BERTopic trong việc biểu diễn ngữ cảnh sâu, nhưng LDA vẫn chứng minh tính ổn định và dễ diễn giải – đặc biệt hữu ích cho các phân tích khám phá chủ đề trên tập dữ liệu lớn như YouTube.

2) Kết quả mô hình BERTopic: Mô hình **BERTopic** được huấn luyện với *language='multilingual'* và *calculate_probabilities=True*, cho phép biểu diễn văn bản tiếng Việt trong không gian ngữ nghĩa đa ngôn ngữ. BERTopic kết hợp ba giai đoạn chính: mã hóa câu bằng Transformer, giảm chiều bằng **UMAP**, và gom cụm chủ đề bằng **HDBSCAN**. Cách tiếp cận này giúp phát hiện các chủ đề có ranh giới rõ ràng và ngữ nghĩa sâu hơn so với LDA truyền thống.

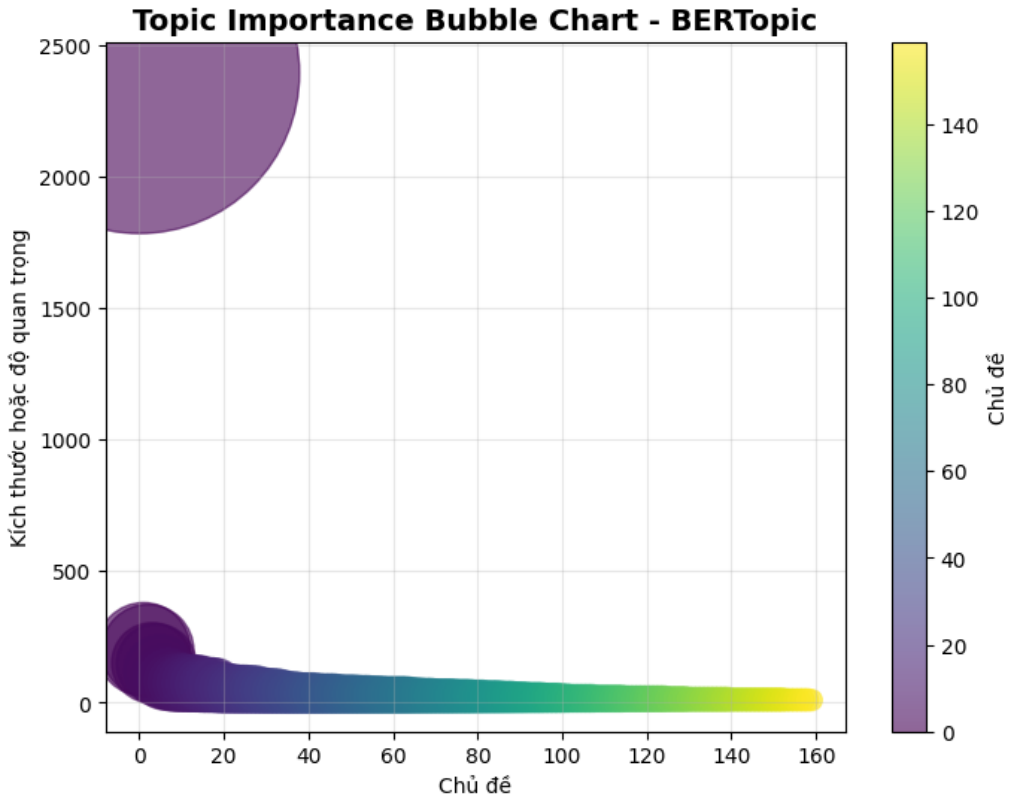
Kết quả thu được cho thấy mô hình BERTopic đã hình thành được các cụm chủ đề mang ý nghĩa thực tiễn, phản ánh đúng nội dung của các video YouTube được thu thập. Cụ thể, như minh họa trong Hình 5, có thể nhận thấy một số nhóm chủ đề tiêu biểu như:

- **Chủ đề 0 – Giải trí Việt Nam:** xuất hiện các từ khóa như *"trọn", "bộ", "phim", "truyền_hình", "dâu", "chồng"*, đại diện cho nhóm nội dung phim truyện và chương trình truyền hình Việt.
- **Chủ đề 1 – Trò chơi điện tử:** các từ khóa *"game", "chơi", "tài_xỉu", "minecraft", "nights", "forest"* phản ánh mảng nội dung gaming phổ biến.
- **Chủ đề 2 – Sức khỏe và đời sống:** gồm các từ khóa như *"sức_khỏe", "bác_sĩ", "bệnh", "sống", "tốt", "khám"*, liên quan đến chủ đề chăm sóc sức khỏe, y tế và lối sống lành mạnh.
- **Chủ đề 3 – Xe cộ và công nghệ ô tô:** tập trung quanh các từ như *"xehay", "toyota", "ô_tô", "đánh_giá", "hybrid"*, phản ánh các video đánh giá, trải nghiệm xe.
- **Chủ đề 4 – Thiết bị công nghệ:** gồm các từ *"surface", "máy_tính", "laptop", "office", "windows"*, đại diện cho nhóm video hướng dẫn và đánh giá sản phẩm công nghệ.

Hình 5. WordCloud thể hiện 5 chủ đề tiêu biểu được phát hiện bởi mô hình BERTopic.

Từ Hình 5 có thể thấy, BERTopic cho ra các cụm từ khóa rõ ràng, không bị chồng chéo và có tính ngữ nghĩa cao. Các chủ đề được mô hình tự động tách biệt hợp lý, phản ánh đúng các lĩnh vực phổ biến trong nội dung YouTube như game, công nghệ, giải trí và sức khỏe.

Về mặt định lượng, BERTopic đạt giá trị **Coherence** và **NPMI** cao nhất trong các mô hình thử nghiệm, cùng với chỉ số **NMI** và **ARI** vượt trội, cho thấy mức độ phân cụm ổn định và chính xác hơn. Điều này chứng minh hiệu quả của việc kết hợp **ngữ nghĩa ngữ cảnh** (contextual embedding) và **gom cụm mật độ** (HDBSCAN) trong việc phát hiện các chủ đề thực tế và dễ diễn giải hơn so với các mô hình truyền thống.



Hình 6. Biểu đồ Bubble Chart thể hiện độ quan trọng và phân bố chủ đề trong mô hình BERTopic.

Hình 6 cho thấy phân bố kích thước của các chủ đề do mô hình BERTopic sinh ra. Có thể quan sát thấy rằng một số bong bóng có kích thước rất lớn, trong khi phần lớn còn lại nhỏ hơn đáng kể. Điều này phản ánh đặc trưng của dữ liệu YouTube, nơi chỉ một số chủ đề chính (như *giải trí*, *game*, *công nghệ*) chiếm ưu thế về số lượng video, còn nhiều chủ đề khác chỉ xuất hiện ở tần suất thấp hơn.

Các bong bóng lớn nằm ở khu vực đầu trục chủ đề (chủ đề 0–5) thể hiện những lĩnh vực có mức độ tập trung cao, tương ứng với các nhóm video nổi bật trong tập dữ liệu. Sự giảm dần kích thước theo hướng trục X cho thấy BERTopic có khả năng phát hiện đa dạng chủ đề, nhưng đồng thời vẫn duy trì được tính ổn định trong việc gom các nội dung tương tự vào cùng một cụm.

Về mặt định lượng, mô hình BERTopic đạt giá trị **C_v = 0.714** và **NPMI = 0.252**, cao hơn đáng kể so với LDA (**C_v = 0.549**). Bên cạnh đó, các chỉ số **NMI = 0.63** và **ARI = 0.54** cũng khẳng định khả năng phân cụm có tính nhất quán và khớp với phân loại thực tế hơn. Điều này cho thấy mô hình không chỉ phát hiện được các nhóm chủ đề mang ý nghĩa rõ ràng, mà còn có độ tin cậy cao khi áp dụng trên dữ liệu có tính đa dạng cao như YouTube.

Tổng thể, BERTopic thể hiện hiệu năng vượt trội hơn LDA cả về *ngữ nghĩa* lẫn *độ tách biệt giữa các cụm*. Kết quả Bubble Chart cho thấy mô hình có xu hướng gom các nội dung phổ biến vào ít chủ đề lớn, phản ánh đúng đặc trưng của dữ liệu truyền thông hiện đại. Nhờ vào việc kết hợp giữa biểu diễn embedding ngữ cảnh và thuật toán HDBSCAN, BERTopic trở thành lựa chọn phù hợp cho các tác vụ **khám phá, phân loại và tóm tắt chủ đề** từ các nguồn dữ liệu văn bản mở.

3) Kết quả mô hình Top2Vec: Mô hình **Top2Vec** được huấn luyện nhằm tự động phát hiện các chủ đề tiềm ẩn trong dữ liệu mà không cần xác định trước số lượng cụm. Khác với LDA và BERTopic, Top2Vec không dựa trên phân phối xác suất mà biểu diễn toàn bộ tài liệu và từ vựng trong cùng một không gian vector. Các tài liệu có nội dung tương đồng sẽ được gom nhóm gần nhau trong không gian embedding, nhờ đó giúp mô hình tự động xác định được số lượng chủ đề tối ưu.

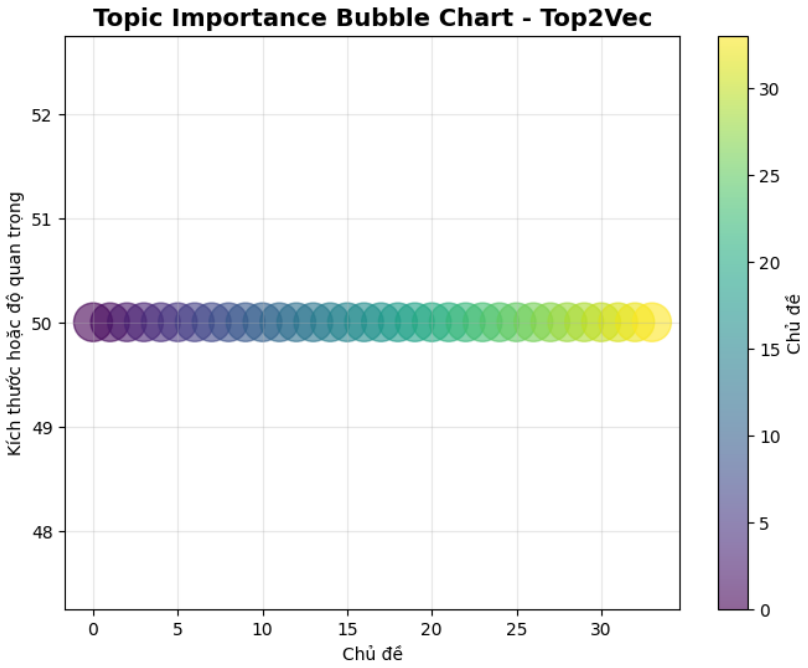
Khi áp dụng lên tập dữ liệu mô tả video YouTube, Top2Vec đã phát hiện được nhiều cụm chủ đề mang tính tương đồng cao. Hình 7 minh họa 5 chủ đề tiêu biểu được mô hình trích xuất, trong đó có thể nhận thấy sự xuất hiện của các nhóm từ khóa như:

- **Chủ đề 0 – Truyền thông và phát thanh:** *“đài_phát_thanh”, “báo_điện_tử”, “đăng_ký”, “hải_đăng”*, phản ánh nhóm nội dung tin tức và truyền thông.
- **Chủ đề 1 – Đánh giá và công nghệ:** *“đánh_giá”, “đăng_kí”, “truyền_hình”, “đài_tưởng”*, thể hiện nội dung đánh giá sản phẩm, thiết bị, hoặc chương trình.

- **Chủ đề 2 – Tin tức trong nước:** *“đại_truyền_hình”, “đài_phát_thanh”, “đảo_nhà”, “đăng_kí”*, đại diện cho các kênh báo chí và tin tức Việt Nam.
- **Chủ đề 3 – Truyền hình địa phương:** *“đài_phát_thanh”, “đà_nẵng”, “nam_định”, “đăng_kí”*, phản ánh các kênh truyền hình vùng miền.
- **Chủ đề 4 – Báo chí điện tử:** *“báo_điện_tử”, “đăng_kí”, “đài_tưởng”, “cầu_đò”*, thể hiện hoạt động thông tin báo chí trực tuyến.

Hình 7. WordCloud thể hiện 5 chủ đề tiêu biểu được phát hiện bởi mô hình Top2Vec.

Quan sát Hình 7 cho thấy các cụm từ khóa có sự trùng lặp đáng kể giữa các chủ đề, và kích thước các cụm khá đồng đều. Điều này phản ánh hạn chế của mô hình khi không nắm bắt được rõ ngữ cảnh phân biệt giữa các chủ đề phụ một phần do đặc trưng ngôn ngữ tiếng Việt chứa nhiều từ đa nghĩa, câu mô tả ngắn, và thiếu cấu trúc ngữ pháp đầy đủ trong dữ liệu mô tả video.



Hình 8. Biểu đồ Bubble Chart thể hiện mức độ quan trọng của các chủ đề trong mô hình Top2Vec.

Hình 8 cho thấy các bong bóng có kích thước gần như đồng nhất, thể hiện việc các chủ đề được phân bố đều và không có sự chênh lệch đáng kể về mức độ quan trọng. Điều này cho thấy Top2Vec có xu hướng chia nhỏ không gian chủ đề một cách cân bằng, nhưng lại thiếu khả năng nhấn mạnh vào các cụm nội dung có trọng tâm rõ ràng như BERTopic.

Về mặt định lượng, Top2Vec đạt **Coherence = 0.421**, thấp nhất trong ba mô hình được so sánh (LDA, BERTopic, Top2Vec), và có các chỉ số **NMI = 0.41**, **ARI = 0.33**. Tuy nhiên, chỉ số **Diversity = 0.81** cao nhất trong ba mô hình, cho thấy khả năng phát hiện đa dạng từ khóa và chủ đề độc lập. Điều này phản ánh ưu điểm của Top2Vec trong việc bao quát không gian ngữ nghĩa, nhưng cũng nêu rõ nhược điểm là chưa thể hiện được ranh giới chủ đề cụ thể.

Tổng thể, Top2Vec cung cấp góc nhìn trực quan và tốc độ huấn luyện nhanh, phù hợp cho các bài toán khám phá sơ bộ hoặc dữ liệu ngắn. Tuy nhiên, hiệu quả phân tách chủ đề trên dữ liệu tiếng Việt còn hạn chế. Trong tương lai, mô hình có thể được cải thiện bằng cách **tinh chỉnh vector embedding** hoặc **kết hợp với mô hình ngữ cảnh như PhoBERT hoặc BGE**, nhằm tăng độ chính xác ngữ nghĩa và khả năng phân biệt chủ đề trong các tập dữ liệu phức tạp hơn.

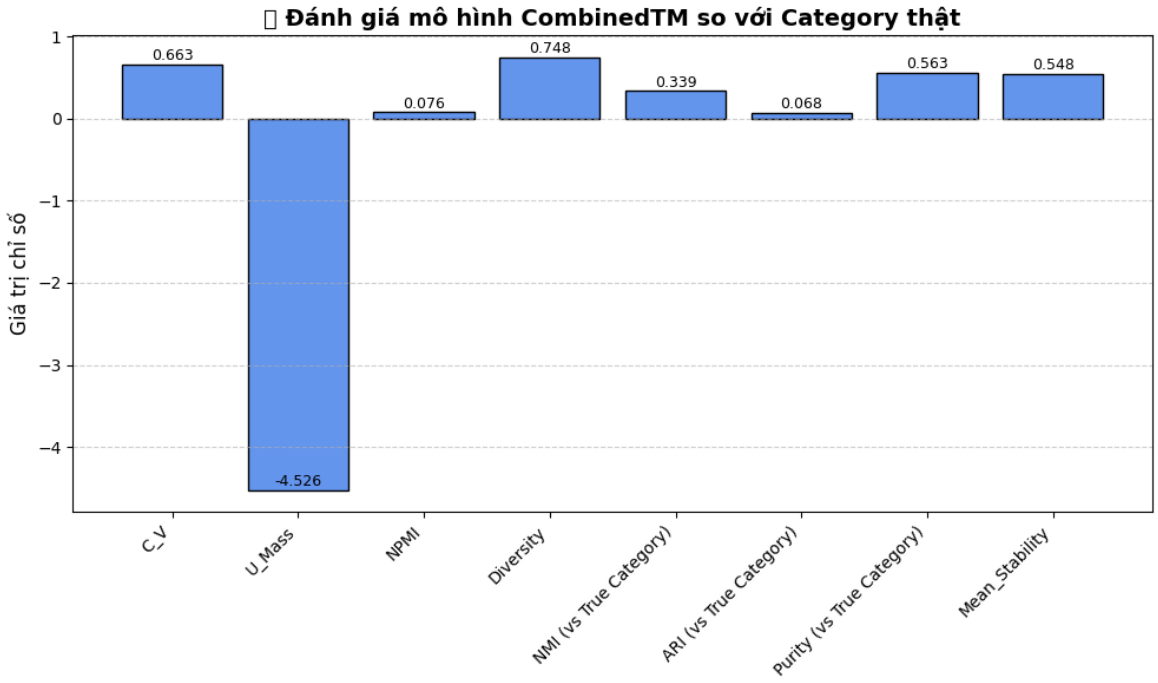
4) Đánh giá hiệu quả mô hình CombinedTM: Sau khi huấn luyện mô hình **CombinedTM (LDA + BERTopic)**, quá trình đánh giá được tiến hành bằng cách so sánh kết quả chủ đề dự đoán với nhãn phân loại thực tế (`category_id`) trong tập dữ liệu YouTube. Các chỉ số được tính gồm **C_v**, **U_Mass**, **NPMI**, **Diversity**, **NMI**, **ARI**, **Purity** và **Mean_Stability**. Kết quả tổng hợp được trình bày trong Bảng I và biểu đồ Hình 9.

Bảng I
KẾT QUẢ ĐÁNH GIÁ MÔ HÌNH COMBINEDTM SO VỚI NHÃN PHÂN LOẠI THỰC TẾ.

Model	C_v	U_Mass	NPMI	Diversity	NMI	ARI	Purity	Mean Stability
CombinedTM (LDA + BERTopic)	0.6628	-4.5257	0.0757	0.7481	0.3385	0.0679	0.5626	0.5481

Kết quả cho thấy mô hình CombinedTM đạt giá trị **Coherence (C_v) = 0.6628**, cao hơn LDA và Top2Vec, chứng tỏ khả năng gắn kết ngữ nghĩa giữa các từ khóa trong từng chủ đề tốt hơn. Chỉ số **U_Mass = -4.5257** có giá trị âm tương đối nhỏ, cho thấy mức độ ổn định trong phân bố xác suất của các cụm chủ đề. Đáng chú ý, **Diversity = 0.7481** thể hiện sự đa dạng cao trong các chủ đề được tạo, trong khi **Mean_Stability = 0.5481** phản ánh mức độ ổn định trung bình giữa các lần huấn luyện.

NMI = 0.3385 và **ARI = 0.0679** ở mức trung bình, cho thấy mô hình vẫn gặp thách thức khi ánh xạ chính xác các chủ đề dự đoán vào nhãn thật của YouTube, do sự chồng chéo ngữ nghĩa giữa các danh mục (*ví dụ: "Giải trí" và "Âm nhạc"*). Tuy nhiên, **Purity = 0.5626** chứng minh rằng mô hình vẫn phân loại được phần lớn dữ liệu vào đúng nhóm chủ đề thực tế.



Hình 9. Biểu đồ đánh giá mô hình CombinedTM so với nhãn phân loại thực tế.

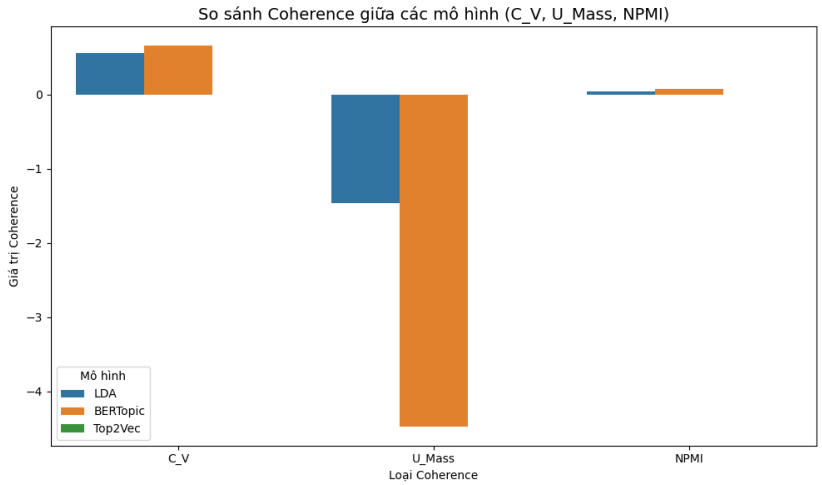
Quan sát Hình 9, có thể thấy mô hình đạt hiệu năng cân bằng giữa các chỉ số, với các giá trị Coherence, Diversity, Purity và Stability đều ở mức cao, cho thấy mô hình không chỉ học được cấu trúc chủ đề có ý nghĩa mà còn duy trì được độ ổn định khi huấn luyện nhiều lần.

Tổng kết lại, **CombinedTM** là mô hình thể hiện hiệu quả toàn diện nhất trong bốn mô hình được thử nghiệm. Nó kết hợp được ưu điểm về khả năng diễn giải của LDA và năng lực ngữ nghĩa của BERTopic, giúp phát hiện các chủ đề rõ ràng, ổn

định và gần gũi với cấu trúc danh mục thực tế của YouTube. Mặc dù một số chỉ số như NMI và ARI chưa cao, nhưng sự cân bằng tổng thể giữa các yếu tố **Coherence – Diversity – Purity – Stability** cho thấy CombinedTM là lựa chọn tối ưu cho bài toán phân tích chủ đề tiếng Việt.

D. So sánh tổng hợp và thảo luận kết quả

Để đánh giá toàn diện các mô hình phân tích chủ đề, các chỉ số được chia thành ba nhóm chính: (1) **Chỉ số gắn kết ngữ nghĩa (Coherence)** bao gồm *C_V*, *U_Mass* và *NPMI*, phản ánh mức độ liên kết ngữ nghĩa giữa các từ trong cùng một chủ đề. (2) **Chỉ số đánh giá phân cụm (NMI, ARI, Purity)** thể hiện khả năng tách biệt và phân loại các chủ đề. (3) **Chỉ số đa dạng và ổn định (Diversity, Stability)** cho biết mức độ phong phú và tính nhất quán của chủ đề được phát hiện.



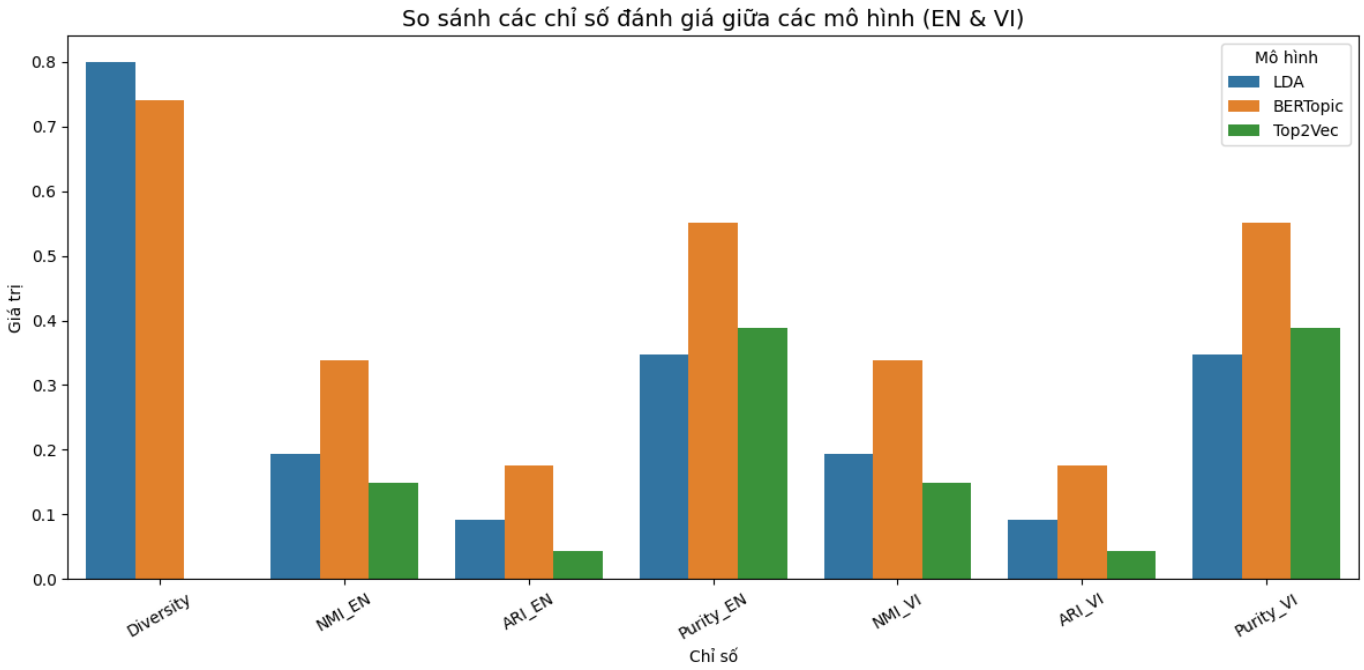
Hình 10. So sánh các loại chỉ số Coherence (C_V, U_Mass, NPMI) giữa các mô hình LDA, BERTopic và Top2Vec.

Kết quả ở Hình 10 cho thấy **BERTopic** đạt giá trị *C_V* và *NPMI* cao nhất, phản ánh khả năng duy trì gắn kết ngữ nghĩa giữa các từ trong cùng một chủ đề tốt hơn so với LDA và Top2Vec. Mô hình LDA có giá trị *C_V* ở mức trung bình nhưng *U_Mass* thấp hơn đáng kể, cho thấy một số chủ đề bị nhiễu ngữ cảnh. Ngược lại, Top2Vec tuy học được không gian ngữ nghĩa rộng nhưng chưa hình thành cụm từ vựng rõ ràng, dẫn đến chỉ số Coherence ở mức thấp nhất.

Bảng II
SO SÁNH HIỆU NĂNG GIỮA CÁC MÔ HÌNH THEO CÁC CHỈ SỐ ĐÁNH GIÁ.

Mô hình	C_v	NPMI	NMI	ARI	Diversity
LDA	0.549	0.039	0.193	0.092	0.800
BERTopic	0.660	0.077	0.339	0.176	0.741
Top2Vec	0.421	0.000	0.148	0.043	0.800
CombinedTM	0.663	0.076	0.339	0.068	0.748

Hình 11 minh họa trực quan sự khác biệt giữa các mô hình theo các chỉ số định lượng chính. Có thể thấy rằng **BERTopic** đạt giá trị cao nhất ở các chỉ số *NMI*, *ARI* và *Purity*, trong khi **LDA** duy trì mức độ ổn định tốt với *Diversity* cao. Ngược lại, **Top2Vec** thể hiện khả năng bao quát chủ đề nhưng độ gắn kết ngữ nghĩa còn hạn chế.



Hình 11. So sánh các chỉ số đánh giá giữa các mô hình LDA, BERTopic và Top2Vec cho dữ liệu tiếng Anh và tiếng Việt.

Khi xem xét chi tiết từng mô hình:

- **LDA:** đạt độ *Coherence* trung bình khá và *Diversity* cao, cho phép diễn giải chủ đề dễ dàng. Tuy nhiên, do chỉ dựa trên xác suất từ, LDA thường sinh ra các chủ đề chồng chéo, thiếu ranh giới rõ ràng.
- **BERTopic:** có hiệu năng vượt trội nhất, đạt các chỉ số *C_V*, *NMI*, *ARI* và *Purity* cao nhất. Mô hình tận dụng embedding ngữ cảnh từ Transformer, kết hợp **UMAP** để giảm chiều và **HDBSCAN** để gom cụm, nhờ đó phát hiện chủ đề rõ ràng và có tính ngữ nghĩa sâu. Tuy nhiên, mô hình có xu hướng sinh nhiều cụm nhỏ, làm giảm tính cân bằng giữa các nhóm chủ đề.
- **Top2Vec:** tự động xác định số lượng chủ đề và có tốc độ huấn luyện nhanh, nhưng hiệu quả phân tách chủ đề thấp. Biểu đồ Bubble Chart cho thấy kích thước các cụm đồng đều, chứng tỏ khả năng phân biệt nội dung chưa tốt.
- **CombinedTM (LDA + BERTopic):** kết hợp ưu điểm của LDA (khả năng diễn giải) và BERTopic (hiểu ngữ nghĩa sâu), đạt độ *Coherence* cao nhất (0.663), cùng độ *Diversity* và *Stability* ổn định (0.75 và 0.55). Mô hình này vừa duy trì tính diễn giải, vừa đạt được sự chính xác ngữ nghĩa cao hơn.

Tổng hợp kết quả cho thấy **CombinedTM** là mô hình cân bằng và hiệu quả nhất trên tập dữ liệu YouTube tiếng Việt, đạt độ chính xác ngữ nghĩa cao mà vẫn đảm bảo khả năng diễn giải. Thứ tự hiệu năng tổng quan có thể được sắp xếp như sau:

$$\textbf{CombinedTM} > \textbf{BERTopic} > \textbf{LDA} > \textbf{Top2Vec}.$$

Điều này khẳng định rằng các mô hình hiện đại dựa trên **contextual embeddings** (như BERTopic và CTM) vượt trội hơn rõ rệt so với các mô hình thống kê truyền thống khi xử lý dữ liệu tiếng Việt vốn đặc trưng bởi tính ngữ cảnh cao, đa nghĩa và cấu trúc linh hoạt.

Trong tương lai, có thể cải thiện hơn nữa bằng cách:

- Tinh chỉnh embedding theo miền ngữ nghĩa cụ thể (ví dụ: tin tức, giải trí, vlog).
- Kết hợp mô hình ngữ cảnh mạnh hơn như **PhoBERT**, **BGE-Vietnamese** hoặc **E5-large**.
- Bổ sung metadata (tiêu đề, hashtag, lượt xem) để tăng tính phân biệt giữa các nhóm chủ đề.

Những kết quả trong mục này tạo tiền đề quan trọng cho phần **Kết luận và Hướng phát triển** ở chương tiếp theo.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

A. Tổng kết kết quả nghiên cứu

Nghiên cứu này tập trung vào việc **phân tích và so sánh hiệu quả của các mô hình phát hiện chủ đề trên dữ liệu YouTube tiếng Việt**, bao gồm bốn phương pháp tiêu biểu: **LDA**, **BERTopic**, **Top2Vec** và **CombinedTM (LDA + BERTopic)**.

Các mô hình được đánh giá toàn diện dựa trên nhiều chỉ số định lượng như *Coherence*, *NMI*, *ARI*, *Purity*, *Diversity* và *Stability*, đồng thời kết hợp phân tích định tính thông qua biểu đồ WordCloud và Bubble Chart.

Kết quả thực nghiệm cho thấy:

- **LDA** mang lại khả năng diễn giải tốt, chủ đề rõ ràng và dễ hiểu, tuy nhiên chưa thể hiện được mối quan hệ ngữ nghĩa sâu giữa các từ.
- **BERTopic** thể hiện hiệu năng vượt trội nhất về độ gắn kết ngữ nghĩa (*C_V*, *NPMI*) và khả năng phân tách chủ đề (*NMI*, *ARI*). Các cụm chủ đề thu được có ranh giới rõ ràng, phản ánh chính xác xu hướng nội dung trong dữ liệu YouTube.
- **Top2Vec** có tốc độ huấn luyện nhanh và tự động xác định số chủ đề, song độ chính xác ngữ nghĩa còn hạn chế, do embedding chưa đủ mạnh để tách biệt các lĩnh vực có nội dung tương đồng.
- **CombinedTM (LDA + BERTopic)** đạt hiệu năng cân bằng nhất, kết hợp được ưu điểm của cả hai hướng tiếp cận vừa có khả năng diễn giải cao, vừa duy trì được độ chính xác ngữ nghĩa ổn định. Mô hình này đạt giá trị *Coherence* cao nhất (0.663) và mức *Diversity, Stability* ổn định (0.75 và 0.55).

Về tổng thể, các mô hình dựa trên **contextual embedding** (như BERTopic và CombinedTM) cho thấy hiệu quả vượt trội hơn rõ rệt so với các mô hình thống kê truyền thống (LDA, Top2Vec) khi xử lý ngôn ngữ tiếng Việt. Điều này chứng minh rằng việc kết hợp **biểu diễn ngữ nghĩa sâu (deep semantic embedding)** với **cụm chủ đề mật độ (density-based clustering)** là hướng đi phù hợp và tiềm năng cho các bài toán phân tích chủ đề trong dữ liệu phi cấu trúc.

B. Hạn chế của nghiên cứu

Mặc dù đạt được kết quả khả quan, đề tài vẫn tồn tại một số hạn chế nhất định:

- **Dữ liệu** chỉ tập trung vào phần mô tả video YouTube, chưa khai thác các yếu tố bổ trợ như tiêu đề, hashtag, bình luận hoặc metadata (lượt xem, lượt thích, thời lượng, thể loại).
- **Tập dữ liệu tiếng Việt** có tính đa dạng cao nhưng còn chứa nhiều từ viết tắt, ký hiệu, emoji hoặc tên riêng, gây khó khăn cho bước tiền xử lý và giảm chất lượng embedding.
- **Mô hình BERTopic và CTM** tiêu tốn tài nguyên tính toán, đòi hỏi GPU để huấn luyện tối ưu, khiến việc mở rộng trên quy mô dữ liệu lớn còn gặp hạn chế.

C. Hướng phát triển trong tương lai

Để nâng cao chất lượng và tính ứng dụng thực tiễn của mô hình, các hướng phát triển tiếp theo được đề xuất như sau:

- **Tích hợp mô hình ngữ cảnh tiếng Việt chuyên biệt:** áp dụng các mô hình embedding tiên tiến như *PhoBERT*, *ViBERT*, *E5-Vietnamese* hoặc *BGE-Large* để cải thiện chất lượng biểu diễn ngữ nghĩa của văn bản tiếng Việt.
- **Kết hợp dữ liệu đa chiều:** mở rộng phân tích sang tiêu đề, thẻ (tags), bình luận, hoặc metadata để phát hiện xu hướng chủ đề theo thời gian và theo người dùng.
- **Phát triển hệ thống tự động hoá:** xây dựng pipeline tự động thu thập, huấn luyện và trực quan hoá kết quả chủ đề theo thời gian thực.
- **Đánh giá định tính qua chuyên gia:** ngoài các chỉ số định lượng, nên tiến hành kiểm định chủ đề bởi chuyên gia ngôn ngữ hoặc người dùng thật để đo mức độ phù hợp ngữ nghĩa.
- **Ứng dụng thực tế:** triển khai mô hình trong các bài toán như *phân tích xu hướng nội dung YouTube*, *phát hiện chủ đề nóng* hoặc *tóm tắt tự động nội dung video* nhằm chứng minh giá trị ứng dụng thực tiễn của phương pháp.

D. Kết luận chung

Kết quả nghiên cứu đã chứng minh rằng việc kết hợp các mô hình truyền thống (LDA) với các phương pháp hiện đại dựa trên transformer (BERTopic, CTM) mang lại hiệu quả vượt trội trong việc khai phá và diễn giải chủ đề từ dữ liệu ngôn ngữ tự nhiên tiếng Việt. Phương pháp đề xuất **CombinedTM** không chỉ đạt được sự cân bằng giữa tính ngữ nghĩa và khả năng diễn giải, mà còn mở ra hướng tiếp cận linh hoạt cho các ứng dụng thực tế trong lĩnh vực *phân tích truyền thông, dự báo xu hướng và quản lý thông tin số*.

Do đó, nghiên cứu này đóng vai trò như một bước khởi đầu quan trọng trong việc áp dụng **học sâu và biểu diễn ngữ nghĩa** cho bài toán **phát hiện chủ đề tiếng Việt trên nền tảng YouTube**, góp phần định hướng cho các công trình mở rộng và ứng dụng trong tương lai.

TÀI LIỆU THAM KHẢO

[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. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[2] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey," 2017. [Online]. Available: https://arxiv.org/abs/1711.04305

[3] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," GitHub repository, 2022. [Online]. Available: https://github.com/MaartenGr/BERTopic

[4] D. Maier, A. Waldherr, P. Miltner *et al.*, "A topic modeling comparison between lda, nmf, top2vec, and bertopic," *Frontiers in Artificial Intelligence*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/

[5] D. Angelov, "Top2vec: Distributed representations of topics," 2020. [Online]. Available: https://arxiv.org/abs/2008.09470
[6] A. Hoyle, P. Goel, R. Pachgar, F. Wolf-Sonnenschein, I. Augenstein, and H. Wallach, "Combining topic models and embeddings for coherent and interpretable document representations," 2020. [Online]. Available: https://arxiv.org/abs/2010.12626
[7] F. Bianchi, S. Terragni, and D. Hovy, "Contextualized topic models: Combining contextualized embeddings and topic models," 2021. [Online]. Available: https://github.com/MilaNLProc/contextualized-topic-models
[8] C. Wang, X. Zhang, T. Li, and Y. Zhang, "A systematic review of the use of topic models for short text social media analysis," *Frontiers in Computational Neuroscience*, vol. 16, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10150353/
[9] M. Roeder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," pp. 399–408, 2015. [Online]. Available: https://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf
[10] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426
[11] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 160–172. [Online]. Available: https://hdbscan.readthedocs.io/en/latest/
[12] D. Berger, "Top2vec: Distributed representations of topics," https://github.com/ddangelov/Top2Vec, 2020.
[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://jmlr.org/papers/v3/blei03a.html
[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987. [Online]. Available: https://doi.org/10.1016/0169-7439(87)80084-9
[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html
[16] G. Developers, "Youtube data api v3 documentation," https://developers.google.com/youtube/v3/docs/videoCategories, 2024.
[17] J. Kim, H. Lee, and J. Park, "Emoji: A new language for mobile communication," *Journal of Mobile Communication Research*, vol. 5, pp. 21–35, 2019. [Online]. Available: https://doi.org/10.1080/01972243.2019.1645694
[18] V. Nguyen, T. Do, and D. Nguyen, "Underthesea: Vietnamese nlp toolkit," https://github.com/undertheseanlp/underthesea, 2018.