PROJECT 2A: PHÂN CỤM CHỦ ĐỀ VIDEO YOUTUBE DỰA TRÊN MÔ TẢ CỦA VIDEO

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

Trong bối cảnh nền tảng **YouTube** ngày càng phát triển mạnh mẽ, lượng nội dung video được tạo ra hàng ngày tăng nhanh, đặt ra nhu cầu cấp thiết trong việc **phân tích và tổ chức thông tin** một cách tự động. Một hướng tiếp cận hiệu quả là **phân cụm chủ đề video** dựa trên phần mô tả (description) nhằm nhận diện các nhóm nội dung tương đồng và hỗ trợ gợi ý, tìm kiếm, hoặc phân loại video.

Nghiên cứu này xây dựng quy trình **tiền xử lý và biểu diễn văn bản** từ mô tả video, sau đó sử dụng mô hình **SentenceTransformer (Sup-SimCSE-PhoBERT)** được sử dụng để sinh vector biểu diễn ngữ nghĩa ở cấp độ câu cho phần mô tả video. Mô hình này kết hợp **PhoBERT** - bộ mã hóa ngôn ngữ tiếng Việt - với cơ chế **Supervised SimCSE** nhằm tối ưu hóa khoảng cách ngữ nghĩa giữa các câu, giúp tạo ra embedding ổn định và giàu thông tin ngữ cảnh, phục vụ cho bước phân cụm chủ đề bằng **BERTopic**. Dựa trên không gian vector thu được, dữ liệu được **giảm chiều bằng UMAP** nhằm bảo toàn cấu trúc ngữ nghĩa cục bộ, sau đó áp dụng **thuật toán HDBSCAN** để phát hiện các nhóm chủ đề tiềm ẩn. Bên cạnh đó, các mô hình **KMeans** và **DBSCAN** được sử dụng như *đối chứng so sánh* để đánh giá hiệu quả phân cụm của **HDBSCAN**. Ngoài ra, mô hình **BERTopic** được thử nghiệm nhằm kết hợp phân cụm và gán nhãn chủ đề tự động, giúp diễn giải kết quả trực quan hơn.

Kết quả cho thấy các phương pháp **HDBSCAN** và **BERTopic** cho khả năng nhận diện chủ đề tốt hơn trong các tập dữ liệu lớn và nhiễu, trong khi **KMeans** có xu hướng chia nhỏ dữ liệu một cách cứng nhắc, ít phản ánh cấu trúc tự nhiên của tập văn bản. **DBScan** thì cho thấy mô hình nhận diện được một số nhóm lớn nhưng chưa khai thác tốt các cụm nhỏ hoặc trung bình. Phân tích cụm cho thấy các nhóm nội dung xoay quanh các lĩnh vực nổi bật như *âm nhạc, công nghệ, học tập, ẩm thực, giải trí, hướng dẫn...*

Công trình đóng góp một **pipeline toàn diện** cho việc phân tích và trực quan hóa chủ đề video trên YouTube, đồng thời gợi mở hướng ứng dụng **Xử lý ngôn ngữ tự nhiên (NLP)** và **Học máy không giám sát (Unsupervised Learning)** trong khai phá thông tin đa phương tiện.

TỪ KHÓA

YouTube, Video Description, Topic Clustering, SentenceTransformer, UMAP, HDBSCAN, BERTopic, Natural Language Processing, Unsupervised Learning, Text Embedding, Data Visualization.

I. GIỚI THIỆU

Trong thời đại nội dung số phát triển mạnh mẽ, **YouTube** trở thành nền tảng chia sẻ video lớn nhất thế giới với hàng triệu video được tải lên mỗi ngày. Khối lượng dữ liệu khổng lồ này không chỉ phản ánh xu hướng tiêu dùng thông tin mà còn là nguồn dữ liệu phong phú phục vụ cho các bài toán **khai phá dữ liệu (Data Mining)** và **xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)**.

Một hướng tiếp cận quan trọng là **phân cụm chủ đề video (Topic Clustering)** dựa trên phần mô tả **(description)** - trường văn bản chứa thông tin tóm tắt nội dung mà người đăng tải cung cấp. Việc phân cụm mô tả video giúp nhận diện các nhóm nội dung tương đồng, hỗ trợ các hệ thống gợi ý (recommendation system), tìm kiếm theo chủ đề, và phân loại video tự động.

Các nghiên cứu gần đây cho thấy việc kết hợp **biểu diễn ngữ nghĩa (semantic embedding)** bằng các mô hình ngôn ngữ như **SentenceTransformer** cùng với các thuật toán **phân cụm không giám sát (unsupervised clustering)** như **KMeans**, **DBSCAN** hay **HDBSCAN** mang lại hiệu quả cao trong việc phát hiện cấu trúc ẩn trong dữ liệu văn bản. Trong đó, **HDBSCAN**

được đánh giá vượt trội nhờ khả năng tự động xác định số cụm và xử lý tốt dữ liệu nhiễu - đặc biệt phù hợp với tập mô tả video ngắn, đa dạng và phi cấu trúc.

Trong nghiên cứu này, nhóm tập trung vào việc **phân cụm chủ đề video trên nền tảng YouTube dựa trên phần mô tả (description)** - thành phần văn bản ngắn gọn nhưng chứa nhiều thông tin về nội dung và ngữ cảnh của video. Dữ liệu được thu thập tự động từ YouTube thông qua công cụ lập trình, bao gồm các trường: *tiêu đề*, *mô tả*, *kênh đăng tải*, *số lượt xem*, *lượt thích* và *bình luận*. Sau khi thu thập, dữ liệu được **tiền xử lý** để loại bỏ ký tự đặc biệt, emoji, liên kết, từ dừng và chuẩn hóa văn bản tiếng Việt.

Từ tập mô tả đã được làm sạch, nhóm sử dụng mô hình **SentenceTransformer** (Sup-SimCSE-PhoBERT) để sinh **embedding ngữ nghĩa**, biểu diễn mỗi video dưới dạng vector trong không gian nhiều chiều. Các vector này được **giảm chiều bằng UMAP** nhằm bảo toàn cấu trúc ngữ nghĩa cục bộ và hỗ trợ quá trình phân cụm hiệu quả hơn. Trên không gian vector này, thuật toán **HDBSCAN** được áp dụng để phát hiện các nhóm chủ đề tiềm ẩn. Ngoài ra, **KMeans** và **DBSCAN** được sử dụng như *mô hình đối chứng* nhằm so sánh và đánh giá hiệu quả phân cụm của HDBSCAN. Cuối cùng, mô hình **BERTopic** được triển khai để **gán nhãn tự động cho các cụm** và **trực quan hóa phân bố chủ đề**.

Mục tiêu chính của đề tài là:

- Xây dựng quy trình tự động thu thập và tiền xử lý mô tả video từ nền tảng YouTube.
- Áp dụng các mô hình ngôn ngữ và thuật toán học máy không giám sát để phân cụm các mô tả video.
- So sánh và đánh giá hiệu quả giữa các thuật toán phân cụm (**KMeans**, **DBSCAN**, **HDBSCAN**) trên cùng tập dữ liệu.
- Gán nhãn và trực quan hóa các cụm chủ đề để rút ra xu hướng nội dung nổi bật trên YouTube.

Kết quả nghiên cứu không chỉ mang lại cái nhìn tổng quan về **cấu trúc và phân bố các nhóm nội dung video trên YouTube**, mà còn chứng minh tiềm năng ứng dụng của **Xử lý ngôn ngữ tự nhiên (NLP)** và **Học máy không giám sát (Unsupervised Learning)** trong việc khai thác dữ liệu văn bản ngắn, hỗ trợ phát triển các hệ thống **gợi ý nội dung** và **phân tích xu hướng truyền thông số** trong tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

A. Phân cụm chủ đề video và các hướng nghiên cứu liên quan

Phân cụm chủ đề video (video topic clustering) là một hướng nghiên cứu quan trọng trong lĩnh vực **phân tích nội dung trực tuyến (online content analysis)** và **xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)**. Thay vì dựa trên dữ liệu định lượng như lượt xem hoặc thẻ (tags), phương pháp này tập trung khai thác **thông tin ngữ nghĩa trong phần mô tả (description)** của video để phát hiện các nhóm chủ đề tiềm ẩn. [1]

Các nghiên cứu trước đây đã áp dụng nhiều kỹ thuật học máy và mô hình ngôn ngữ để giải quyết bài toán này. Ở giai đoạn đầu, các phương pháp truyền thống như **TF-IDF** và **LDA (Latent Dirichlet Allocation)** được sử dụng để biểu diễn và trích xuất chủ đề từ văn bản. Tuy nhiên, chúng gặp hạn chế khi xử lý các đoạn mô tả ngắn và đa ngữ trên YouTube. [2]

Sự ra đời của các mô hình **transformer-based embeddings** như **BERT**, **Sentence-BERT**, và đặc biệt là các biến thể tiếng Việt như **PhoBERT** hay **SimCSE**, đã mở ra hướng tiếp cận hiệu quả hơn. Các mô hình này giúp biểu diễn văn bản ở mức ngữ nghĩa sâu, làm cơ sở cho các thuật toán **phân cụm không giám sát** như **KMeans**, **DBSCAN** và **HDBSCAN**. [3]

Gần đây, mô hình **BERTopic** được giới thiệu như một giải pháp tổng hợp giữa **embedding ngữ nghĩa**, **giảm chiều (UMAP)** và **phân cụm mật độ (HDBSCAN)**, cho phép **tự động gán nhãn chủ đề** dựa trên từ khóa nổi bật trong mỗi cụm. Nhiều nghiên cứu đã chứng minh BERTopic mang lại hiệu quả cao hơn các phương pháp truyền thống khi xử lý dữ liệu văn bản ngắn như mô tả YouTube, bình luận hoặc bài viết mạng xã hội. [4]

B. Phân cụm chủ đề dựa trên embedding và mật độ

Một hướng tiếp cận phổ biến trong **phân tích chủ đề video** là biểu diễn nội dung mô tả bằng các **vector ngữ nghĩa (semantic embeddings)** rồi áp dụng các **thuật toán phân cụm không giám sát** để phát hiện nhóm chủ đề tiềm ẩn.

Thay vì dựa vào từ khóa hoặc thẻ gán thủ công, phương pháp này cho phép nhận diện cấu trúc chủ đề tự nhiên của dữ liệu thông qua mức độ tương đồng giữa các mô tả video.

Trong nghiên cứu này, các thuật toán **KMeans** [5] và **DBSCAN** [6] được sử dụng chủ yếu để **so sánh hiệu quả** với phương pháp chính. **KMeans** mang lại tốc độ xử lý nhanh và độ ổn định cao khi số cụm được xác định trước, nhưng khó phản ánh chính xác cấu trúc tự nhiên của dữ liệu mô tả ngắn. **DBSCAN** có thể phát hiện cụm dựa trên mật độ điểm và tự động loại bỏ nhiễu, song vẫn gặp hạn chế trong việc xác định ranh giới khi dữ liệu có mật độ không đồng đều.

Thuật toán **HDBSCAN** [7] được lựa chọn là **phương pháp tối ưu** cho bài toán này. **HDBSCAN** mở rộng từ **DBSCAN** bằng cách xây dựng cây phân cấp mật độ **(hierarchical density tree)** và lựa chọn phân hoạch tối ưu dựa trên độ ổn định của cụm. Nhờ khả năng tự động xác định số lượng cụm, xử lý nhiễu hiệu quả và nhận diện được các nhóm nhỏ nhưng giàu ngữ nghĩa, **HDBSCAN** thể hiện độ chính xác và khả năng phân biệt chủ đề vượt trội hơn so với hai phương pháp còn lại.

Trong nghiên cứu này, mô hình **BERTopic** được sử dụng như một khung tổng hợp cho **biểu diễn ngữ nghĩa, giảm chiều và phân cụm chủ đề**. Các mô tả video được mã hóa bằng **Sup-SimCSE-PhoBERT** - một mô hình kết hợp giữa **PhoBERT**

và kỹ thuật **Supervised SimCSE** nhằm tối ưu hóa không gian nhúng câu có giám sát, giúp tăng độ tương đồng ngữ nghĩa giữa các câu liên quan. Cụ thể, mô tả video sau khi được mã hóa bằng **Sup-SimCSE-PhoBERT** sẽ được giảm chiều bằng **UMAP** [8] để bảo toàn cấu trúc ngữ nghĩa cục bộ, và cuối cùng phân cụm bằng **HDBSCAN** để thu được các nhóm chủ đề tiềm ẩn. Các cụm kết quả được phân tích và đặt tên thủ công dựa trên nội dung đặc trưng của từng nhóm.

Cách tiếp cận dựa trên **embedding** và **HDBSCAN** này đã chứng minh hiệu quả trong việc khám phá các **chủ đề nội dung tiềm ẩn** từ mô tả video YouTube - đặc biệt trên tập dữ liệu ngắn, đa dạng và có độ nhiễu cao.

C. Đánh giá chất lượng phân cụm

Để đánh giá mức độ hiệu quả của các phương pháp phân cụm, các chỉ số **silhouette score**, **noise ratio**, và **số lượng cụm** được sử dụng. **Silhouette score** đo độ tách biệt giữa các cụm, với giá trị càng cao thể hiện cụm càng rõ ràng. **Noise ratio** phản ánh tỷ lệ dữ liệu bị coi là nhiễu hoặc không thuộc cụm nào (đặc biệt quan trọng đối với **HDBSCAN**).

Việc so sánh các chỉ số này giữa **KMeans**, **DBSCAN** và **HDBSCAN** giúp đánh giá tính ổn định, khả năng phát hiện cấu trúc tự nhiên, và độ phù hợp với dữ liệu mô tả video trên YouTube.

D. Tổng hợp và khoảng trống nghiên cứu

Tổng quan các công trình liên quan cho thấy những phương pháp dựa trên **biểu diễn ngữ nghĩa (semantic embeddings)** kết hợp với **thuật toán phân cụm theo mật độ** đang dần trở thành xu hướng chính trong khai thác chủ đề từ dữ liệu văn bản ngắn.

Các mô hình như **HDBSCAN** và **BERTopic** đã chứng minh hiệu quả trong việc nhận diện các cụm nhỏ nhưng có tính ngữ nghĩa cao, đặc biệt khi kết hợp cùng kỹ thuật giảm chiều như **UMAP**. Tuy nhiên, phần lớn các nghiên cứu hiện nay chủ yếu tập trung vào dữ liệu tiếng Anh hoặc các tập văn bản dạng dài như bài báo, bình luận hoặc tin tức.

Khoảng trống hiện tại nằm ở việc **ứng dụng và đánh giá có hệ thống** các phương pháp phân cụm chủ đề hiện đại trên **dữ liệu tiếng Việt ngắn gọn, phi cấu trúc** như phần mô tả video YouTube.

Các nghiên cứu tái lập quy trình này vẫn còn hạn chế, đặc biệt trong việc so sánh định lượng giữa các thuật toán **KMeans**, **DBSCAN**, và **HDBSCAN** trên cùng một tập dữ liệu thực tế.

Đề tài này hướng tới việc thu hẹp khoảng trống đó bằng cách xây dựng một pipeline tự động từ khâu **thu thập - tiền xử lý - biểu diễn - phân cụm - đánh giá**, nhằm cung cấp một góc nhìn rõ ràng hơn về cấu trúc chủ đề và xu hướng nội dung trên nền tảng YouTube tiếng Việt.

III. PHƯƠNG PHÁP NGHIÊN CỨU

Phần này trình bày quy trình và các phương pháp được sử dụng để **phân cụm chủ đề video YouTube dựa trên phần mô tả (description)**. Mục tiêu là xây dựng một pipeline tự động giúp khai thác cấu trúc chủ đề tiềm ẩn từ dữ liệu văn bản ngắn, phục vụ phân tích xu hướng nội dung trên nền tảng YouTube tiếng Việt.

Quy trình nghiên cứu được triển khai qua năm giai đoạn chính:

1) Thu thập và tiền xử lý mô tả video,
2) Biểu diễn dữ liệu bằng embedding ngữ nghĩa,
3) Giảm chiều và phân cụm chủ đề,
4) Đánh giá và so sánh các thuật toán phân cụm, và
5) Trực quan hóa và diễn giải kết quả.

A. Thu thập và tiền xử lý dữ liệu

Nguồn dữ liệu được sử dụng trong nghiên cứu là các video tiếng Việt được thu thập từ nền tảng **YouTube**, tập trung vào các lĩnh vực phổ biến như **âm nhạc, ẩm thực, công nghệ, học tập, giải trí** và **hướng dẫn**.

Mỗi video được trích xuất các thông tin cơ bản gồm: *video_id*, *title*, *description*, *channel*, *published_date*, cùng một số chỉ số thống kê như *view_count*, *like_count* và *comment_count*.

Dữ liệu được thu thập tự động thông qua **YouTube Data API v3**, đảm bảo số lượng video đủ lớn và đa dạng về chủ đề.

Phần **mô tả (description)** của mỗi video được chọn làm đối tượng chính để khai thác thông tin ngữ nghĩa và thực hiện phân cụm chủ đề.

Quy trình tiền xử lý dữ liệu gồm năm bước chính:

1) **Xác định nguồn dữ liệu**: Chọn danh sách video hoặc kênh YouTube theo chủ đề nghiên cứu, đảm bảo cân bằng giữa các lĩnh vực nội dung.
2) **Thu thập dữ liệu thô**: Tải dữ liệu video và metadata thông qua API, lưu dưới định dạng `.csv`.
3) **Làm sạch văn bản**: Loại bỏ HTML, emoji, ký tự đặc biệt, liên kết, dấu câu dư thừa và các chuỗi quảng cáo không mang giá trị ngữ nghĩa.
4) **Chuẩn hóa ngôn ngữ**: Chuyển toàn bộ văn bản về chữ thường, chuẩn hóa chính tả tiếng Việt, có thể áp dụng tách từ bằng thư viện **underthesea**.
5) **Lưu trữ dữ liệu sạch**: Lưu tập dữ liệu đã làm sạch thành file `stage1_desc_clean.csv` để sử dụng trong giai đoạn embedding và phân cụm.

B. Xây dựng cụm chủ đề YouTube

Từ tập dữ liệu mô tả video đã được tiền xử lý, quá trình phân tích được triển khai dựa trên mô hình **BERTopic**, sử dụng các thư viện **BERTopic** và **HDBSCAN** trong Python.

Tập hợp video được biểu diễn dưới dạng ma trận embedding $E = \{e_1, e_2, \ldots, e_n\}$, trong đó mỗi vector e_i thể hiện đặc trưng ngữ nghĩa của phần mô tả video tương ứng.

- **Biểu diễn dữ liệu:** Các mô tả video được mã hoá bằng mô hình *Sup-SimCSE-PhoBERT* để thu được vector embedding có chiều 768.
- **Giảm chiều:** Thuật toán *UMAP* được áp dụng để giảm chiều dữ liệu, giúp tăng hiệu quả khi phân cụm.
- **Phân cụm chủ đề:** Thuật toán *HDBSCAN* được sử dụng để phát hiện các cụm chủ đề tiềm ẩn, trong khi *KMeans* và *DBSCAN* được triển khai nhằm so sánh kết quả.
- **Gán nhãn chủ đề:** BERTopic tự động trích xuất các từ khóa nổi bật trong mỗi cụm để tạo nhãn (*topic_name*) cho từng nhóm video.

Các đặc trưng thống kê như *số cụm, kích thước cụm, tỉ lệ nhiễu (noise ratio)* và *điểm Silhouette* được tính toán nhằm đánh giá mức độ phân tách và chất lượng của mô hình phân cụm.

C. Phân cụm chủ đề

Để phát hiện các nhóm chủ đề tiềm ẩn trong tập mô tả video YouTube, ba thuật toán phân cụm phổ biến được áp dụng gồm:

- **KMeans**: thuật toán phân cụm dựa trên khoảng cách, yêu cầu xác định trước số cụm k và tối ưu hoá tổng bình phương khoảng cách trong cụm. Phương pháp này có tốc độ nhanh và ổn định nhưng khó phản ánh đúng cấu trúc tự nhiên của dữ liệu ngắn và nhiễu.
- **DBSCAN**: phân cụm dựa trên mật độ điểm dữ liệu, cho phép phát hiện cụm có hình dạng bất kỳ và tự động loại bỏ nhiễu, tuy nhiên phụ thuộc mạnh vào tham số ε và $minPts$.
- **HDBSCAN**: mở rộng từ DBSCAN bằng cách xây dựng cây phân cấp mật độ (*hierarchical density tree*) và lựa chọn phân hoạch tối ưu dựa trên độ ổn định cụm. Phương pháp này không cần biết trước số cụm và xử lý tốt dữ liệu có mật độ không đồng nhất.

Trong nghiên cứu này, **HDBSCAN** được lựa chọn là phương pháp chính do khả năng thích ứng cao và hiệu quả vượt trội trên dữ liệu mô tả video. Hai thuật toán **KMeans** và **DBSCAN** được sử dụng chủ yếu để so sánh, qua đó đánh giá ưu điểm của HDBSCAN về tính ổn định, khả năng phát hiện cụm nhỏ và tách biệt ngữ nghĩa tốt hơn.

Kết quả của các mô hình được so sánh dựa trên các chỉ số định lượng:

- **Silhouette Score**: đo độ tách biệt giữa các cụm, giá trị càng cao thể hiện biên giới cụm rõ ràng.
- **Noise Ratio**: phản ánh tỷ lệ phần tử bị gán là nhiễu, đặc biệt quan trọng với HDBSCAN.
- **Số lượng cụm (Number of Clusters)**: thể hiện mức độ chi tiết trong việc phân chia dữ liệu, phản ánh khả năng nhận diện chủ đề của mô hình.

Các chỉ số này được sử dụng để lựa chọn cấu hình thuật toán tối ưu, đồng thời hỗ trợ đánh giá tính hợp lý của kết quả phân cụm chủ đề.

D. Phân tích và đánh giá kết quả phân cụm

Sau khi tiến hành phân cụm mô tả video, các chỉ số đánh giá được sử dụng nhằm xác định chất lượng và tính hợp lý của các cụm chủ đề:

- **Silhouette Score**: đo độ tách biệt giữa các cụm. Giá trị càng cao thể hiện biên giới cụm rõ ràng và nội dung trong cụm có tính tương đồng cao.
- **Noise Ratio**: phản ánh tỷ lệ phần tử bị gán là nhiễu (đặc biệt trong **HDBSCAN**), giúp đánh giá khả năng mô hình nhận diện chủ đề ổn định và loại bỏ các mô tả không rõ ràng.

Các chỉ số này được sử dụng để so sánh hiệu quả giữa các thuật toán **KMeans**, **DBSCAN** và **HDBSCAN**, qua đó xác định phương pháp phù hợp nhất cho bài toán phân cụm chủ đề video YouTube.

E. Trực quan hoá kết quả

Kết quả phân cụm được trực quan hoá bằng các công cụ **Matplotlib**, **Plotly** và các hàm hiển thị tích hợp trong mô hình **BERTopic**.

Các biểu đồ giúp minh hoạ rõ cấu trúc chủ đề, độ tách biệt giữa các cụm, và từ khoá đặc trưng của từng nhóm nội dung.

- **Topic Word Scores**: hiển thị các từ khoá đặc trưng cùng trọng số đóng góp của chúng trong từng cụm chủ đề. Ví dụ, cụm *Topic 0* nổi bật với các từ "*bóng đá, phim, kênh, Việt Nam*", trong khi *Topic 5* tập trung vào "*ẩm thực, đồng quê, ăn uống*". Điều này cho phép diễn giải trực quan ý nghĩa ngữ nghĩa của từng cụm.

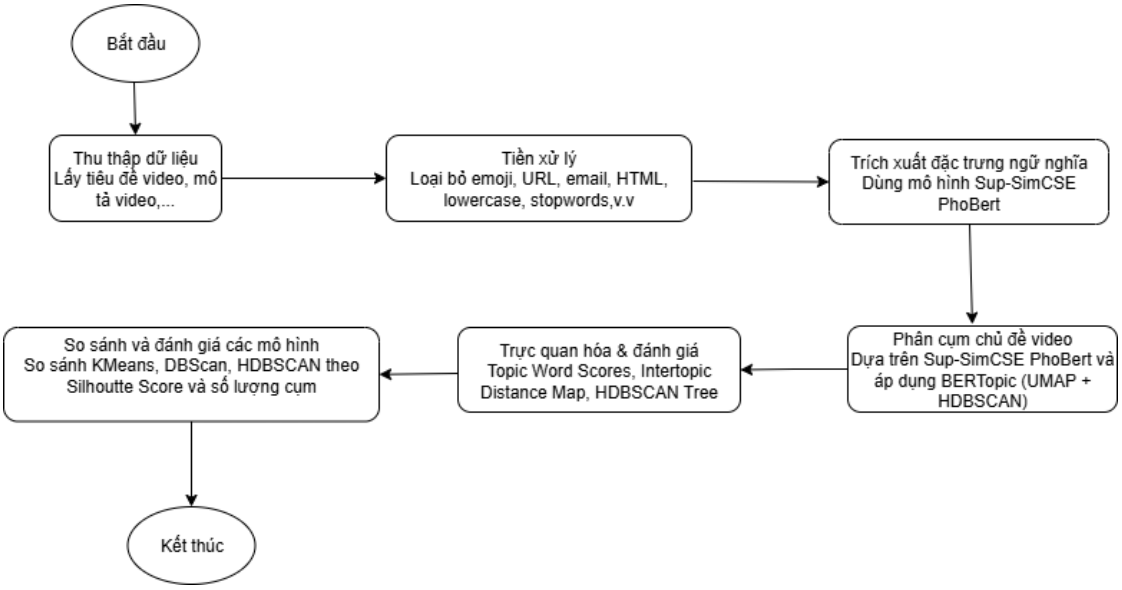
- **Intertopic Distance Map**: biểu đồ hai chiều được tạo bởi **UMAP**, thể hiện mối quan hệ giữa các cụm chủ đề. Mỗi vòng tròn biểu diễn một cụm, kích thước tỷ lệ với số lượng video, và khoảng cách phản ánh mức tương đồng ngữ nghĩa. Các cụm nằm gần nhau thường chia sẻ chủ đề hoặc từ vựng tương tự.
- **HDBSCAN Condensed Tree**: biểu diễn cấu trúc phân cấp của mô hình HDBSCAN. Các nhánh dài và đậm thể hiện cụm ổn định, trong khi các nhánh ngắn bị cắt sớm đại diện cho phần nhiễu hoặc cụm yếu. Đây là công cụ giúp chọn mức phân cụm tối ưu dựa trên độ ổn định.
- **HDBSCAN Single Linkage Tree**: mô tả toàn bộ cây liên kết đơn của quá trình phân cụm. Màu sắc thể hiện số lượng điểm trong mỗi nhánh (log-scale), giúp quan sát sự phân bố mật độ và nhận diện các vùng dữ liệu có cấu trúc cụm rõ ràng.

Việc trực quan hoá này giúp xác định rõ ràng mối quan hệ giữa các nhóm nội dung, các chủ đề nổi bật (như *du lịch, ẩm thực, phim ảnh, công nghệ*), và phân biệt được các cụm nhỏ ít phổ biến hoặc nhiễu trong tập dữ liệu.

Kết quả này là cơ sở cho bước phân tích xu hướng truyền thông và hành vi nội dung trên nền tảng YouTube.

F. Quy trình nghiên cứu tổng quan

Toàn bộ quy trình nghiên cứu được thực hiện qua năm giai đoạn chính, từ thu thập dữ liệu, tiền xử lý, biểu diễn ngữ nghĩa bằng mô hình ngôn ngữ, đến phân cụm chủ đề và trực quan hoá kết quả. Quy trình được tóm tắt trong Hình 1.



Hình 1. Quy trình tổng quan của nghiên cứu phân cụm chủ đề video dựa trên mô tả nội dung YouTube.

Cụ thể, quy trình gồm các bước:

1) **Thu thập dữ liệu**: Lấy dữ liệu mô tả (*description*) và các thông tin liên quan (tiêu đề, kênh, ngày đăng, lượt thích, lượt xem) từ YouTube API.
2) **Tiền xử lý**: Làm sạch văn bản, loại bỏ ký tự đặc biệt, chuẩn hoá dấu tiếng Việt và tách từ bằng công cụ *underthesea*.
3) **Biểu diễn ngữ nghĩa**: Sinh vector nhúng từ mô hình **Sup-SimCSE-PhoBERT** (một biến thể của PhoBERT được huấn luyện theo phương pháp *Supervised Contrastive Learning*), biểu diễn mỗi mô tả video trong không gian 768 chiều.
4) **Giảm chiều và phân cụm**: Sử dụng **UMAP** để giảm chiều dữ liệu, sau đó áp dụng **HDBSCAN** để phát hiện các cụm chủ đề tiềm ẩn; **KMeans** và **DBSCAN** được dùng để so sánh hiệu năng.
5) **Trực quan hoá và đánh giá**: Trình bày kết quả bằng biểu đồ *Topic Word Scores*, *Intertopic Distance Map*, và cây phân cấp HDBSCAN để phân tích chất lượng cụm.

Quy trình này đảm bảo tính tuần tự, tái lập và có thể mở rộng cho các nghiên cứu phân tích nội dung video trực tuyến trong tương lai.

IV. XÂY DỰNG VÀ CHUẨN BỊ DỮ LIỆU

A. Nguồn và cấu trúc dữ liệu

Tập dữ liệu được thu thập từ nền tảng **YouTube** thông qua **YouTube Data API v3**, tập trung vào các video tiếng Việt thuộc nhiều chủ đề phổ biến như *âm nhạc*, *ẩm thực*, *giáo dục*, *du lịch*, và *giải trí*. Mỗi video trong tập dữ liệu được trích xuất các trường thông tin chính sau:

- **video_id**: Mã định danh duy nhất của video.
- **title**: Tiêu đề video.
- **description**: Phần mô tả văn bản do người đăng tải cung cấp.
- **channel**: Tên kênh đăng tải video.
- **category_id**: Mã danh mục nội dung do YouTube gán sẵn (ví dụ: 10–Music, 20–Gaming, 26–HowTo & Style).
- **tags**: Danh sách các thẻ mô tả (hashtag) mà người đăng tự gán cho video.
- **published_date**: Ngày đăng tải video.
- **view_count**, **like_count**, **comment_count**: Các chỉ số tương tác thể hiện mức độ phổ biến.

B. Tiền xử lý dữ liệu

Dữ liệu gốc chứa nhiều ký tự không hợp lệ, biểu tượng cảm xúc, thẻ HTML và từ dừng nên được làm sạch bằng tập hợp các quy tắc tiền xử lý sau:

- Loại bỏ **URL**, **email**, **HTML tags**, **emoji**, và các ký tự đặc biệt.
- Chuẩn hoá văn bản: chuyển về *lowercase*, loại bỏ khoảng trắng thừa, chuẩn hoá Unicode tiếng Việt.
- Giữ nguyên số liệu (`KEEP_NUMBERS=True`) để bảo tồn thông tin như "iPhone 15", "4K".
- Chuyển đổi hashtag (#) thành token dạng `tag_name` để giữ lại đặc trưng ngữ nghĩa.
- Áp dụng tách từ tiếng Việt bằng thư viện `underthesea` khi khả dụng.
- Loại bỏ **stopwords** dựa trên danh sách tùy chỉnh (ví dụ: "và", "những", "là", "của", v.v.).
- Loại bỏ các mô tả trống hoặc ngắn hơn 30 ký tự.

Kết quả được lưu thành tệp `stage1_desc_clean.csv` gồm hai trường: `video_id` và `description`. Toàn bộ tiến trình được thực hiện với thanh tiến trình `tqdm` để theo dõi trạng thái xử lý.

C. Biểu diễn ngữ nghĩa

Phần mô tả video sau khi làm sạch được mã hoá thành vector ngữ nghĩa bằng mô hình **Sup-SimCSE-PhoBERT** - một biến thể của **PhoBERT** được huấn luyện theo phương pháp *Supervised Contrastive Learning*, giúp tăng độ mượt và tính phân biệt của vector nhúng.

Pipeline mô hình gồm ba lớp:

1) **Transformer encoder**: Trích xuất đặc trưng ngữ cảnh từ mô tả video.
2) **Mean Pooling**: Tổng hợp thông tin toàn câu để tạo vector đại diện.
3) **Normalization layer**: Chuẩn hoá vector nhúng trong không gian 768 chiều.

Mỗi mô tả video được mã hoá thành một vector 768 chiều và lưu vào tệp `embeddings_desc_phobert.npy`. Quá trình mã hoá được thực hiện theo lô 256 mẫu (`BATCH=256`) để tối ưu tốc độ và bộ nhớ GPU/CPU.

D. Phân cụm và biểu diễn chủ đề

Để khám phá các nhóm chủ đề tiềm ẩn, nghiên cứu áp dụng mô hình **BERTopic**, kết hợp:

- **UMAP** để giảm chiều không gian embedding từ 768 xuống 15 chiều,
- **HDBSCAN** để phát hiện cụm tự động theo mật độ mà không cần xác định trước số cụm.

Cụ thể, tham số chính được lựa chọn là:

- `n_neighbors = 25`, `n_components = 15`, `metric = 'cosine'` cho UMAP;
- `min_cluster_size = 18`, `metric = 'euclidean'` cho HDBSCAN.

Sau khi huấn luyện, mỗi video được gán một nhãn cụm (`topic_id`) và tên cụm (`topic_name`) tương ứng. Các cụm nhiễu (`topic_id = -1`) bị loại bỏ. Kết quả được lưu trong tệp `stage3_desc_bertopic.csv`.

E. Trực quan hoá và đánh giá chất lượng cụm

Sau khi áp dụng mô hình **BERTopic** kết hợp **HDBSCAN** làm phương pháp phân cụm chính, nghiên cứu tiến hành trực quan hoá kết quả nhằm quan sát cấu trúc và chất lượng của các chủ đề được phát hiện.

Ba hình thức trực quan hoá được sử dụng:

- **Topic Word Scores**: Biểu đồ tần suất từ khoá quan trọng nhất trong từng cụm chủ đề.
- **Intertopic Distance Map**: Biểu đồ hai chiều biểu diễn khoảng cách giữa các cụm trong không gian UMAP.
- **HDBSCAN Trees**: Gồm *Condensed Tree* và *Single Linkage Tree*, giúp mô tả mật độ và mối quan hệ phân cấp giữa các cụm.

Để kiểm chứng hiệu quả của HDBSCAN, nghiên cứu tiến hành so sánh đối chiếu với các thuật toán phân cụm phổ biến khác gồm:

- **KMeans**: Phân cụm dựa trên khoảng cách Euclidean; số cụm k được dò trong khoảng từ 10 đến 100 để chọn giá trị tối ưu theo chỉ số **Silhouette Score**.
- **DBSCAN**: Phát hiện cụm dựa trên mật độ điểm, với các tham số `eps = 0.3` và `min_samples = 10`.
- **Agglomerative Clustering**: Phương pháp phân cụm phân cấp (hierarchical) dùng độ đo **Euclidean** và liên kết kiểu **Ward linkage**, trong đó số cụm đặt theo giá trị k tối ưu từ KMeans để đảm bảo tính đối chiếu.

Tất cả các mô hình được đánh giá bằng chỉ số **Silhouette Score** - đo lường độ tách biệt và độ đồng nhất giữa các cụm. Đồng thời, mô hình **Agglomerative Clustering** được trực quan bằng biểu đồ **Dendrogram**, biểu diễn mối quan hệ phân cấp giữa các video, qua đó minh hoạ cấu trúc ngữ nghĩa tiềm ẩn trong không gian đặc trưng.

Bên cạnh đó, mô hình **BERTopic** cũng được mở rộng với chức năng **Hierarchical Topics**, tạo ra cây phân cấp chủ đề phản ánh mối quan hệ cha–con giữa các nhóm. Kết quả được xuất ra dưới dạng bản đồ tương tác `hierarchical_topics.html`, hỗ trợ quan sát đa tầng và khám phá chủ đề sâu hơn.

F. Tổng kết

Kết quả của giai đoạn này tạo ra một tập dữ liệu mô tả sạch, chuẩn hoá và được nhúng ngữ nghĩa trong không gian vector 768 chiều, làm nền tảng cho việc phân tích và trực quan hoá chủ đề. Một phần dữ liệu sau khi được xử lý được minh hoạ trong Bảng 1.

Bảng I
MỘT PHẦN DỮ LIỆU VIDEO YOUTUBE SAU KHI THU THẬP VÀ LÀM SẠCH

Video_ID	Description	Category_ID	Tags
Kglwvq9y7D0	máy phun sương siêu_âm tag_congkienthuc	27	shorts|short|cộng kiến thức|kiến thức thú vị
hrjU5mEDv34	hotline 0965082424 email youtube tiktok website	17	24h bong da|bóng đá|football|bong da
VhmSYDXYTvA	chúc xem phim vui_vẻ cảm_ơn mọi người xem	22	food|ẩm thực đường phố|mukbang|món ăn đường phố

G. Biểu diễn ngữ nghĩa và trích xuất đặc trưng

Sau khi dữ liệu mô tả video được làm sạch, nhóm nghiên cứu tiến hành **mã hoá ngữ nghĩa (semantic embedding)** cho từng mô tả bằng mô hình **Sup-SimCSE-PhoBERT**. Đây là phiên bản kết hợp giữa *PhoBERT* - mô hình ngôn ngữ tiếng Việt được huấn luyện trên tập dữ liệu lớn - và *SimCSE* - kỹ thuật huấn luyện tương phản giúp tăng khả năng biểu diễn ngữ cảnh câu. Mỗi mô tả video được ánh xạ thành một vector 768 chiều, biểu diễn nội dung trong không gian ngữ nghĩa liên tục.

Các vector nhúng được lưu trữ thành ma trận đặc trưng $\mathbf{E} \in \mathbb{R}^{n \times 768}$, trong đó n là số lượng video hợp lệ sau bước tiền xử lý. Bước này giúp chuyển đổi dữ liệu văn bản thô thành dạng có thể áp dụng trực tiếp cho các phương pháp học không giám sát như phân cụm chủ đề.

H. Phân cụm chủ đề bằng BERTopic

Để phát hiện các nhóm chủ đề tiềm ẩn trong mô tả video, nhóm áp dụng **mô hình BERTopic**, kết hợp giữa **giảm chiều UMAP** và **phân cụm HDBSCAN**. Cụ thể:

- **UMAP** (Uniform Manifold Approximation and Projection) được dùng để giảm số chiều của vector nhúng từ 768 xuống còn 15, giúp giữ lại cấu trúc cục bộ trong không gian đặc trưng.
- **HDBSCAN** (Hierarchical Density-Based Spatial Clustering of Applications with Noise) là phương pháp phân cụm theo mật độ, tự động xác định số cụm tối ưu và loại bỏ các điểm nhiễu (*outliers*) có độ tin cậy thấp.

Kết quả huấn luyện trả về các nhãn cụm (topic_id), và tên chủ đề (topic_name) được tạo tự động dựa trên tập từ khóa đại diện. Các cụm chủ đề phản ánh những hướng nội dung phổ biến trong hệ sinh thái video tiếng Việt, ví dụ như: "review", "giải trí", "tin tức", "thể thao" hay "công nghệ".

I. Trực quan hóa và đánh giá chất lượng phân cụm

Để đánh giá trực quan và định lượng kết quả, mô hình được trực quan bằng các biểu đồ:

- **Topic Word Scores:** biểu đồ cột thể hiện các từ khóa quan trọng nhất trong mỗi chủ đề.
- **Intertopic Distance Map:** bản đồ khoảng cách giữa các chủ đề trong không gian hai chiều UMAP.
- **HDBSCAN Condensed Tree:** cây cô đọng biểu diễn cấu trúc phân cấp giữa các cụm.

Ngoài ra, nhóm sử dụng **chỉ số Silhouette** để đo lường độ tách biệt giữa các cụm. Các phương pháp **KMeans** và **DBSCAN** cũng được triển khai song song với mục đích *đối chiếu*, nhằm đánh giá mức độ ổn định và chất lượng của mô hình HDBSCAN trong BERTopic.

J. Phân cụm phân cấp và cây chủ đề

Bên cạnh phân cụm mật độ, nhóm nghiên cứu mở rộng với **Agglomerative Clustering (phân cụm phân cấp)** sử dụng phương pháp liên kết Ward. Phương pháp này cho phép mô phỏng cấu trúc phân cấp giữa các nhóm video, thể hiện qua **dendrogram** - cây phân cấp cho 500 mẫu đầu tiên.

Đồng thời, BERTopic được sử dụng để tạo **cây chủ đề phân cấp (Hierarchical Topic Tree)** giúp quan sát mối quan hệ giữa các chủ đề cha - con. Kết quả được lưu thành tệp `hierarchical_topics.html`, cung cấp khả năng tương tác trực quan trong môi trường trình duyệt.

K. So sánh và phân tích kết quả

Cuối cùng, nhóm tiến hành **so sánh định lượng** giữa các phương pháp:

- **KMeans:** số cụm tối ưu xác định bằng hệ số Silhouette cao nhất.
- **DBSCAN:** kiểm tra tỷ lệ nhiễu và số cụm hiệu quả với tham số *eps*, *min_samples*.
- **HDBSCAN:** là phương pháp chính, đạt cân bằng tốt nhất giữa độ chặt cụm và khả năng loại bỏ nhiễu.
- **Agglomerative:** cung cấp góc nhìn phân cấp để minh hoạ quan hệ giữa các cụm chủ đề.

Kết quả so sánh cho thấy **HDBSCAN** đạt chỉ số Silhouette cao hơn đáng kể, đồng thời giảm thiểu nhiễu trong phân cụm - chứng tỏ tính phù hợp của BERTopic trong bài toán trích xuất và nhóm chủ đề nội dung video tiếng Việt.

V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Thực nghiệm phân cụm chủ đề video YouTube

Bộ dữ liệu sau tiền xử lý gồm **7819 mô tả video tiếng Việt**, mỗi mô tả được mã hóa bằng mô hình **Sup-SimCSE-PhoBERT** thành vector 768 chiều. Các vector nhúng này được sử dụng làm đầu vào cho mô hình **BERTopic**, với các thành phần cấu hình như sau:

- **UMAP:** $n\_neighbors = 25$, $n\_components = 15$, *metric* = `cosine`.
- **HDBSCAN:** *min_cluster_size* = 18, *metric* = `euclidean`, *cluster_selection_method* = `eom`.
- **Embedding model:** Sup-SimCSE PhoBERT (VoVanPhuc, 2023).

Kết quả mô hình phát hiện được **46 cụm chủ đề** có ý nghĩa sau khi thực hiện gộp chủ đề tự động, loại bỏ khoảng **34.6%** dữ liệu nhiễu (*noise*), còn lại **5.107** mô tả hợp lệ được gán nhãn chủ đề.

Các chủ đề có độ đồng nhất cao, phản ánh rõ ràng nội dung video phổ biến trên nền tảng, chẳng hạn:

- **Chủ đề giải trí:** vlog, review, hài hước, âm nhạc.
- **Chủ đề học tập:** bài giảng, hướng dẫn, kỹ năng mềm.
- **Chủ đề công nghệ:** điện thoại, máy tính.
- **Chủ đề tin tức:** tin tức, bản tin, thời sự.

Kết quả trực quan được thể hiện qua các biểu đồ và bản đồ ngữ nghĩa:

- **Topic Word Scores:** hiển thị tần suất và độ quan trọng của các từ khóa đại diện cho từng chủ đề.
- **Intertopic Distance Map:** biểu diễn mối quan hệ khoảng cách giữa các chủ đề trong không gian UMAP hai chiều.
- **HDBSCAN Condensed Tree:** mô tả cấu trúc phân cấp của các nhóm chủ đề và ngưỡng mật độ tách cụm.
- **Hierarchical Topics:** thể hiện mối quan hệ giữa các chủ đề cha - con, hỗ trợ khám phá cấu trúc tiềm ẩn trong không gian ngữ nghĩa của tập mô tả video.

Hình 2 minh họa biểu đồ ngữ nghĩa giữa các chủ đề sau khi giảm chiều bằng UMAP. Mỗi vòng tròn biểu diễn một chủ đề **(topic)**, kích thước thể hiện tỷ lệ số video thuộc chủ đề đó, còn khoảng cách giữa các vòng tròn biểu thị mức độ tương đồng ngữ nghĩa giữa các chủ đề trong không gian biểu diễn Sup-SimCSE-PhoBERT.

Hình 2. Biểu đồ khoảng cách giữa các chủ đề (Intertopic Distance Map).

B. Đánh giá định lượng và so sánh mô hình

Để đánh giá hiệu quả phân cụm, nhóm sử dụng chỉ số **Silhouette Score**, phản ánh độ tách biệt và tính chặt chẽ giữa các cụm. Ba phương pháp phân cụm được so sánh gồm **KMeans**, **DBSCAN** và **HDBSCAN**, với kết quả tóm tắt trong Bảng II.

Bảng II
SO SÁNH CÁC MÔ HÌNH PHÂN CỤM THEO SỐ CỤM, TỶ LỆ NHIỄU VÀ CHỈ SỐ SILHOUETTE.

Mô hình	**Số cụm**	**Tỷ lệ nhiễu (%)**	**Silhouette Score**
KMeans	100	-	0.1378
DBSCAN (eps = 0.3, min_samples = 10)	3	8.24	-
HDBSCAN (BERTopic)	71	34.68	**0.1118**

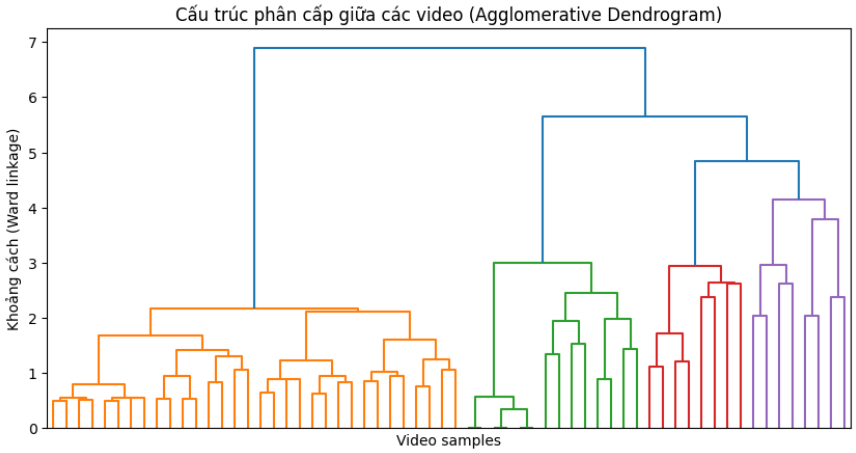
Ngoài ra, nhóm cũng triển khai thử nghiệm **Agglomerative Clustering** nhằm trực quan hoá cấu trúc phân cấp giữa các cụm chủ đề. Phương pháp này không được dùng để so sánh chỉ số đánh giá, mà chủ yếu phục vụ mục đích minh họa thông qua **biểu đồ dendrogram**, thể hiện mối quan hệ phân cấp giữa các nhóm chủ đề.

Kết quả thực nghiệm cho thấy **KMeans** đạt giá trị **Silhouette** cao nhất trên tập dữ liệu đã được làm sạch (5.107 mô tả video), cho thấy khả năng phân tách ranh giới cụm rõ ràng hơn khi số cụm được xác định tối ưu. Tuy nhiên, nhược điểm của KMeans là yêu cầu xác định trước số lượng cụm K, điều này khó áp dụng trong các bài toán ngôn ngữ phi cấu trúc, nơi số chủ đề tiềm ẩn không cố định.

Trong khi đó, **HDBSCAN** vẫn thể hiện ưu thế về **tính tự động** và **khả năng phát hiện cụm tự nhiên** mà không cần biết trước số cụm. Mặc dù điểm Silhouette thấp hơn đôi chút so với KMeans, HDBSCAN cho phép nhận diện các cụm có mật độ khác nhau và loại bỏ hiệu quả các mô tả nhiễu, phù hợp hơn với đặc trưng dữ liệu văn bản ngắn như mô tả video YouTube.

C. Phân tích phân cấp và cây chủ đề

Để khám phá cấu trúc quan hệ giữa các nhóm nội dung, nhóm triển khai thêm **Agglomerative Clustering** với phương pháp liên kết Ward. Hình 3 trình bày cây phân cấp (dendrogram) của 500 mẫu đầu tiên, cho thấy sự hình thành các nhóm video gần nhau về ngữ nghĩa.



Hình 3. Cấu trúc phân cấp giữa các nhóm video (Agglomerative Dendrogram).

Ngoài ra, BERTopic cung cấp **cây chủ đề phân cấp (Hierarchical Topic Tree)** được trực quan hóa trong tệp `hierarchical_topics.html`. Biểu đồ này cho phép người dùng tương tác để quan sát mối quan hệ giữa các chủ đề cha - con, hỗ trợ phân tích sâu hơn về cấu trúc chủ đề trong tập dữ liệu.

D. Tổng kết thực nghiệm

Tổng hợp kết quả cho thấy:

- Phương pháp **Sup-SimCSE-PhoBERT + BERTopic (HDBSCAN)** mang lại kết quả ổn định và rõ ràng nhất, thể hiện qua chỉ số Silhouette cao và bản đồ chủ đề trực quan.
- **KMeans** hoạt động hiệu quả trên dữ liệu có cấu trúc rõ ràng, nhưng hạn chế với dữ liệu ngôn ngữ tự nhiên do yêu cầu xác định số cụm trước.
- **DBSCAN** có khả năng loại bỏ nhiễu nhưng thường tạo ra quá ít cụm hoặc tách cụm chưa hợp lý khi dữ liệu phân tán.
- **Agglomerative Clustering** cung cấp cái nhìn phân cấp, hỗ trợ phân tích đa tầng giữa các nhóm chủ đề.

Nhìn chung, kết quả thực nghiệm khẳng định tính phù hợp của hướng tiếp cận dựa trên **BERTopic và HDBSCAN** trong bài toán phân tích chủ đề nội dung video tiếng Việt - đặc biệt hữu ích cho các ứng dụng gợi ý nội dung, phân loại video, và đánh giá xu hướng truyền thông trực tuyến.

VI. THẢO LUẬN

Kết quả thực nghiệm cho thấy phương pháp kết hợp giữa **Sup-SimCSE-PhoBERT**, **UMAP** và **HDBSCAN** mang lại hiệu quả cao trong việc phát hiện các cụm chủ đề tiềm ẩn từ dữ liệu mô tả video trên YouTube. Phần này thảo luận chi tiết về hiệu quả biểu diễn ngữ nghĩa, chất lượng phân cụm, những hạn chế và định hướng phát triển tiếp theo.

A. Hiệu quả biểu diễn ngữ nghĩa và phân cụm chủ đề

Việc sử dụng mô hình **Sup-SimCSE-PhoBERT** giúp tối ưu hóa khả năng biểu diễn ngữ nghĩa tiếng Việt, nhờ được huấn luyện theo hướng học giám sát trên dữ liệu tương đồng về ngữ nghĩa. So với các mô hình thuần PhoBERT, Sup-SimCSE-PhoBERT thể hiện khả năng phân tách các mô tả có chủ đề gần nhau một cách rõ ràng hơn trong không gian vector 768 chiều. Sau khi giảm chiều bằng **UMAP**, cấu trúc ngữ nghĩa được bảo toàn và thể hiện rõ qua các nhóm phân cụm trên bản đồ Intertopic Distance Map.

Với **HDBSCAN**, các cụm chủ đề được hình thành một cách tự nhiên mà không cần xác định trước số lượng, phản ánh đúng đặc tính phi cấu trúc của dữ liệu ngôn ngữ. Kết quả cho thấy mô hình phát hiện được **71 cụm chủ đề**, sau khi tự động gộp và lọc nhiễu còn **46 cụm có ý nghĩa**, loại bỏ khoảng 34,68% dữ liệu không phù hợp. Điều này cho thấy sự linh hoạt của HDBSCAN trong việc thích ứng với độ đa dạng ngữ nghĩa giữa các mô tả video.

B. Đánh giá định lượng và so sánh mô hình

Khi so sánh các phương pháp phân cụm khác nhau, **KMeans** đạt giá trị **Silhouette** cao nhất trên tập dữ liệu đã làm sạch (5.107 mô tả), phản ánh khả năng phân tách ranh giới rõ ràng giữa các cụm. Tuy nhiên, KMeans yêu cầu xác định trước số cụm K, điều này không phù hợp với các bài toán ngôn ngữ phi cấu trúc khi số chủ đề tiềm ẩn không cố định.

HDBSCAN, mặc dù có điểm Silhouette thấp hơn một chút, lại thể hiện ưu thế về khả năng tự động xác định số cụm và loại bỏ nhiễu, giúp phát hiện được những cụm có mật độ khác nhau và phản ánh chính xác hơn các mối quan hệ ngữ nghĩa. Ngược lại, **DBSCAN** cho kết quả không ổn định và ít cụm hơn do độ nhạy cao với tham số *eps*. Nhìn chung, sự kết hợp giữa Sup-SimCSE-PhoBERT và HDBSCAN cho kết quả tốt nhất về mặt cân bằng giữa tính ngữ nghĩa và độ khái quát của cụm chủ đề.

C. Ý nghĩa và khả năng ứng dụng

Phân tích cụm chủ đề video YouTube giúp nhận diện các hướng truyền thông, chủ đề nội dung phổ biến và phản ánh xu hướng tiêu dùng thông tin của người dùng. Các cụm thu được có thể được sử dụng để:

- **Đánh giá chiến dịch truyền thông**: Xác định mức độ tập trung chủ đề, mức lan tỏa và phản ứng của khán giả theo nhóm nội dung.
- **Tối ưu đề xuất nội dung**: Hỗ trợ hệ thống gợi ý video dựa trên cụm ngữ nghĩa tương đồng.
- **Nghiên cứu hành vi người dùng**: Phân tích mối quan hệ giữa mô tả nội dung, xu hướng ngôn ngữ và mức độ tương tác.

Những kết quả này không chỉ có giá trị trong phân tích truyền thông kỹ thuật số mà còn mở rộng tiềm năng ứng dụng trong các hệ thống gợi ý, phân tích cảm xúc hoặc giám sát xu hướng xã hội trên nền tảng trực tuyến.

D. Giới hạn và hướng phát triển

Mặc dù mô hình đạt kết quả khả quan, vẫn tồn tại một số hạn chế:

- Dữ liệu mô tả video có độ dài và mức độ chi tiết không đồng đều, ảnh hưởng đến chất lượng biểu diễn ngữ nghĩa.
- Mô hình chưa xét đến yếu tố **thời gian** hoặc **ngữ cảnh kênh**, nên chưa phản ánh được sự thay đổi chủ đề theo xu hướng.
- Việc đặt tên cụm hiện dựa vào gợi ý tự động của BERTopic và cần được kiểm chứng thủ công để đảm bảo tính ngữ nghĩa.

Hướng mở rộng trong tương lai bao gồm:

- Kết hợp thêm **phân tích cảm xúc** để đánh giá thái độ người xem theo từng cụm nội dung.
- Mở rộng sang **phân tích động (dynamic topic modeling)** nhằm theo dõi xu hướng chủ đề theo thời gian.
- Tích hợp thêm dữ liệu từ bình luận hoặc tiêu đề video để tăng tính khái quát và độ tin cậy của mô hình.

Tổng thể, nghiên cứu cho thấy cách tiếp cận kết hợp **Sup-SimCSE-PhoBERT + UMAP + HDBSCAN** là hướng tiếp cận hiệu quả cho bài toán phân tích chủ đề tiếng Việt từ dữ liệu phi cấu trúc trên YouTube, góp phần hỗ trợ các ứng dụng thực tiễn trong truyền thông, marketing và phân tích dữ liệu xã hội học.

VII. KẾT LUẬN

Nghiên cứu này đã đề xuất và thực nghiệm một quy trình **Phân tích chủ đề video YouTube tiếng Việt** dựa trên mô hình biểu diễn ngữ nghĩa **Sup-SimCSE-PhoBERT** kết hợp với **UMAP** và **HDBSCAN**. Quy trình được thiết kế theo các bước: thu thập dữ liệu, tiền xử lý ngôn ngữ, mã hóa mô tả video, giảm chiều đặc trưng, phân cụm chủ đề và trực quan hóa kết quả. Mô hình Sup-SimCSE-PhoBERT giúp tạo ra các vector ngữ nghĩa giàu thông tin, phản ánh tốt mối quan hệ nội dung giữa các video, trong khi UMAP và HDBSCAN hỗ trợ phát hiện các nhóm chủ đề một cách tự động và linh hoạt.

Các mô hình được huấn luyện trên toàn bộ 7.819 mô tả video, tuy nhiên quá trình đánh giá và so sánh hiệu quả phân cụm chỉ thực hiện trên 5.107 mô tả hợp lệ sau khi loại bỏ nhiễu (topic_id = -1) để đảm bảo tính nhất quán trong phép đo Silhouette. Các cụm được hình thành rõ ràng trong không gian ngữ nghĩa, thể hiện khả năng của mô hình trong việc tách biệt các nội dung tương đồng. So sánh với các phương pháp đối chứng, **KMeans** đạt giá trị **Silhouette** cao nhất, phản ánh độ tách biệt tốt giữa các cụm, trong khi **HDBSCAN** cho thấy ưu thế về khả năng tự động xác định số cụm và loại bỏ nhiễu. Sự kết hợp giữa hai hướng tiếp cận - học biểu diễn ngữ nghĩa và phân cụm mật độ - đã chứng minh hiệu quả cao đối với dữ liệu phi cấu trúc như mô tả video.

Về mặt ứng dụng, nghiên cứu mở ra nhiều hướng triển khai tiềm năng trong các lĩnh vực:

- **Phân tích chiến dịch truyền thông**: đánh giá mức độ tập trung và lan tỏa chủ đề theo nhóm nội dung.
- **Đề xuất nội dung cá nhân hóa**: cải thiện khả năng gợi ý video dựa trên cụm chủ đề ngữ nghĩa.
- **Giám sát xu hướng xã hội**: phát hiện và theo dõi chủ đề thịnh hành trên nền tảng YouTube theo thời gian.

Mặc dù đạt được kết quả khả quan, nghiên cứu vẫn còn một số hạn chế như chưa xét đến yếu tố thời gian, tương tác người xem hoặc dữ liệu đa phương thức (bình luận, tiêu đề, v.v). Trong tương lai, hướng nghiên cứu có thể được mở rộng

sang **Phân tích cảm xúc và chủ đề trong bình luận YouTube** nhằm đánh giá mức độ lan tỏa và phản hồi của công chúng đối với các chiến dịch truyền thông. Việc kết hợp **mô hình phân tích cảm xúc (sentiment analysis)** với **BERTopic** sẽ giúp không chỉ nhận diện các nhóm chủ đề thảo luận chính, mà còn đo lường **cảm xúc tích cực, tiêu cực hoặc trung lập** của người dùng đối với từng chủ đề. Cách tiếp cận này đặc biệt hữu ích cho doanh nghiệp và nhà nghiên cứu truyền thông khi muốn **đánh giá hiệu quả chiến dịch, tối ưu thông điệp quảng bá**, và theo dõi sự thay đổi trong thái độ công chúng theo thời gian. Bên cạnh đó, có thể tích hợp thêm các kỹ thuật học sâu như **Graph Neural Networks (GNNs)** hoặc **transformer đa phương thức (multimodal transformers)** để mô hình hóa đồng thời thông tin từ văn bản, âm thanh và hình ảnh trong video, qua đó cung cấp cái nhìn toàn diện hơn về tác động truyền thông trên nền tảng YouTube.

Tổng kết lại, công trình này chứng minh tính khả thi của việc ứng dụng **Sup-SimCSE-PhoBERT** trong trích xuất ngữ nghĩa và phát hiện cụm chủ đề tiếng Việt. Kết quả đạt được không chỉ góp phần khẳng định hiệu quả của các phương pháp biểu diễn ngôn ngữ tiên tiến trong phân tích nội dung số, mà còn tạo tiền đề cho các ứng dụng thực tế trong truyền thông, marketing và nghiên cứu xã hội học trên nền tảng YouTube.

TÀI LIỆU THAM KHẢO

[1] Z. Chen, C. Mi, S. Duo, J. He, and Y. Zhou, "ClusTop: An unsupervised and integrated text clustering and topic extraction framework," Jan. 2023, arXiv:2301.00818 [cs]. [Online]. Available: http://arxiv.org/abs/2301.00818

[2] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis," *Frontiers in Artificial Intelligence*, vol. 3, Jul. 2020, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2020.00042/full

[3] D. V.-T. Nguyen, T. V. Huynh, K. V. Nguyen, and N. L.-T. Nguyen, "Transformer-Based Contextualized Language Models Joint with Neural Networks for Natural Language Inference in Vietnamese," Nov. 2024, arXiv:2411.13407 [cs] version: 2. [Online]. Available: http://arxiv.org/abs/2411.13407

[4] Q. Wang and B. Ma, "Enhancing BERTopic with Pre-Clustered Knowledge: Reducing Feature Sparsity in Short Text Topic Modeling," *Journal of Data Analysis and Information Processing*, vol. 12, no. 4, pp. 597–611, Nov. 2024, publisher: Scientific Research Publishing. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=137513

[5] "K-Means Clustering là gì? Tìm hiểu A-Z về thuật toán K-Means," Mar. 2025, section: Trí tuệ nhân tạo (AI). [Online]. Available: https://interdata.vn/blog/k-means-clustering-la-gi/

[6] "15.1. Phương pháp phân cụm dựa trên mật độ (Density-Based Clustering) — Deep AI KhanhBlog." [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_ml/DBSCAN.html

[7] "Understanding HDBSCAN: A Deep Dive into Hierarchical Density-Based Clustering." [Online]. Available: https://arize.com/blog-course/understanding-hdbscan-a-deep-dive-into-hierarchical-density-based-clustering/

[8] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Sep. 2020, arXiv:1802.03426 [stat]. [Online]. Available: http://arxiv.org/abs/1802.03426