NGHIÊN CỨU XÂY DỰNG BỘ DỮ LIỆU PHỤ ĐỀ VIDEO HỌC LIỆU NGÀNH CÔNG NGHỆ THÔNG TIN

Trần Thị Thu Phương, Nguyễn Quốc Tuấn, Lê Thị Hằng
Trường Đại học Thủ đô Hà Nội

Tóm tắt: *Nghiên cứu này tập trung vào việc xây dựng một bộ dữ liệu phụ đề video chuyên ngành CNTT nhằm nâng cao khả năng tiếp cận học liệu và hỗ trợ phát triển các ứng dụng NLP trong giáo dục. Nghiên cứu đề xuất một quy trình thu thập và xử lý dữ liệu bài bản, từ lựa chọn nguồn, trích xuất, làm sạch, chuẩn hóa đến kiểm định chất lượng dữ liệu. Kết quả là một bộ dữ liệu có tính học thuật cao, làm nền tảng cho các nghiên cứu và ứng dụng trong lĩnh vực giáo dục CNTT*

Từ khóa: *phụ đề học liệu; xử lý ngôn ngữ tự nhiên; trí tuệ nhân tạo trong giáo dục; học liệu số ngành CNTT.*

Nhận bài ngày 20.04.2025; gửi phản biện, chỉnh sửa, duyệt đăng ngày 30.05.2025
Liên hệ tác giả: Nguyễn Quốc Tuấn; email: nqtuan@daihocthudo.edu.vn

1. ĐẶT VẤN ĐỀ

Trong bối cảnh thế giới đang bước vào cuộc cách mạng công nghiệp lần thứ tư, dữ liệu đã trở thành nền tảng cho đổi mới sáng tạo trong nhiều lĩnh vực, đặc biệt là giáo dục và Công nghệ Thông tin (CNTT). Sự bùng nổ của dữ liệu trên toàn cầu, với ước tính đạt 175 zettabyte vào năm 2025 [1], đặt ra yêu cầu cấp thiết về việc tổ chức và khai thác dữ liệu hiệu quả trong giáo dục đại học. Trong đó, các tài liệu học liệu kỹ thuật số, như video giảng dạy và khóa học trực tuyến, ngày càng đóng vai trò quan trọng trong quá trình đào tạo.

Tuy nhiên, việc thiếu các bộ dữ liệu phụ đề học thuật chất lượng cao và mang tính chuyên ngành đang là một thách thức lớn, đặc biệt đối với các video giáo dục trong lĩnh vực CNTT. Nghiên cứu của World Economic Forum (2020) chỉ ra rằng, mặc dù các kỹ năng CNTT như AI, học máy và lập trình là những năng lực quan trọng trong tương lai, nhiều người học vẫn gặp rào cản về ngôn ngữ và khả năng tiếp cận học liệu số [2]. Điều này càng cho thấy sự cần thiết của các giải pháp hỗ trợ như phụ đề, dịch thuật và chatbot học tập để tăng tính cá nhân hóa và khả năng tiếp cận nội dung số, một xu hướng được 74% các cơ sở giáo dục đại học quan tâm (EDUCAUSE, 2022).[3]

Tại Việt Nam, nhu cầu xây dựng cơ sở dữ liệu học liệu số chuyên ngành CNTT, bao gồm cả phụ đề video, đang tăng nhanh cùng với quá trình số hóa của các trường đại học. Tuy nhiên, việc đảm bảo chất lượng của các dữ liệu này vẫn còn là một vấn đề nan giải do thiếu các quy trình chuẩn hóa và bộ tiêu chí toàn diện, gây khó khăn cho việc ứng dụng AI và các công nghệ Xử lý ngôn ngữ tự nhiên (NLP) trong giáo dục.

Để giải quyết những thách thức trên, nghiên cứu này đặt mục tiêu xây dựng một bộ dữ liệu phụ đề video chuyên ngành CNTT, dựa trên các tiêu chuẩn học thuật rõ ràng và có tính ứng dụng thực tiễn cao trong giáo dục đại học. Nghiên cứu không chỉ tập trung vào việc thu thập và xử lý dữ liệu mà còn hướng đến việc đánh giá khả năng ứng dụng của các mô hình tóm tắt, dịch thuật và trích xuất thông tin, những công nghệ có tiềm năng định hình tương lai của giáo dục số. Hướng tiếp cận này phù hợp với xu thế phát triển của các nền tảng học trực tuyến hàng đầu như Coursera, edX và Udacity, nơi mà việc tích hợp phụ đề đa ngôn ngữ và

các công cụ hỗ trợ truy cập ngày càng được chú trọng để nâng cao trải nghiệm học tập. Các nghiên cứu trong lĩnh vực giáo dục trực tuyến đã chứng minh rằng, phụ đề chính xác và hỗ trợ ngôn ngữ không chỉ cải thiện khả năng tiếp thu kiến thức mà còn tăng cường tương tác và tỷ lệ hoàn thành khóa học, đặc biệt đối với người học không sử dụng ngôn ngữ bản địa hoặc có nhu cầu hỗ trợ đặc biệt.

Tóm lại, việc xây dựng dữ liệu phụ đề học liệu CNTT có vai trò quan trọng trong việc nâng cao chất lượng đào tạo và tạo điều kiện thuận lợi cho việc tích hợp AI vào giáo dục một cách toàn diện và bền vững.

2. NỘI DUNG

2.1. Tổng quan về xử lý phụ đề và chuẩn hóa dữ liệu video

2.1.1. Khái niệm và vai trò của phụ đề

Phụ đề là phần văn bản hiển thị đồng bộ với nội dung âm thanh của video, thường thể hiện lời thoại, nội dung thuyết minh hoặc mô tả âm thanh. Trong bối cảnh giáo dục số, đặc biệt là các video học liệu chuyên ngành, phụ đề không chỉ hỗ trợ người học tiếp cận tốt hơn nội dung bài giảng mà còn đóng vai trò quan trọng trong việc:[4]

- Hỗ trợ người học có nhu cầu đặc biệt (ví dụ: người khiếm thính),
- Tăng khả năng tiếp thu nội dung trong môi trường đa nhiệm (xem video không bật âm thanh),
- Tạo nền tảng cho các ứng dụng học máy như tóm tắt nội dung, phân tích dữ liệu ngôn ngữ, dịch tự động, tìm kiếm nội dung theo văn bản...

2.1.2. Các loại phụ đề và định dạng kỹ thuật

Phụ đề có thể được chia thành:

- Phụ đề cứng (open captions): gắn liền với video, không thể tắt hoặc chỉnh sửa.
- Phụ đề mềm (closed captions): tồn tại dưới dạng file riêng biệt, có thể bật/tắt hoặc xử lý lại.

Các định dạng phổ biến trong phụ đề học liệu gồm:

- SRT (SubRip Subtitle): đơn giản, dễ sử dụng, phù hợp cho nghiên cứu NLP.
- VTT (WebVTT): thường dùng cho nền tảng web, hỗ trợ thêm siêu dữ liệu.
- ASS/SSA: hỗ trợ định dạng phức tạp, chủ yếu dùng trong phụ đề phim.

Việc chọn định dạng phụ đề cần phù hợp với mục tiêu xử lý dữ liệu, đảm bảo khả năng đồng bộ, dễ trích xuất và tích hợp với các mô hình AI.

2.1.3. Quy trình xử lý và tạo phụ đề

- Quá trình tạo phụ đề có thể thực hiện theo ba cách:
- Thủ công: nghe – gõ lại lời thoại – căn thời gian. Tuy chính xác nhưng mất thời gian.
- Bán tự động: sử dụng hệ thống nhận dạng giọng nói (như Whisper hoặc Google ASR) để tạo bản nháp, sau đó con người hiệu đính và căn chỉnh lại thời gian. Đây là hướng tiếp cận được sử dụng trong nghiên cứu này để cân bằng giữa độ chính xác và hiệu suất.
- Tự động hoàn toàn: sử dụng các hệ thống AI tạo và căn chỉnh phụ đề, nhưng độ chính xác còn phụ thuộc nhiều vào âm thanh, ngữ cảnh và từ vựng chuyên ngành.

2.1.4. Chuẩn hóa dữ liệu

Để phục vụ mục tiêu nghiên cứu và ứng dụng AI, việc chuẩn hóa dữ liệu là bước then chốt. Các tiêu chí chuẩn hóa bao gồm:

- Định dạng thống nhất: sử dụng định dạng phụ đề (SRT/VTT), định dạng video (MP4).
- Cấu trúc metadata rõ ràng: mỗi mẫu dữ liệu cần gắn thông tin về tiêu đề, lĩnh vực, giảng viên, thời lượng, ngôn ngữ, nguồn gốc...

- Từ vựng chuyên ngành được xử lý nhất quán: chuẩn hóa thuật ngữ, xử lý viết tắt, đồng bộ giữa các ngôn ngữ.
- Thời gian hiển thị phụ đề chính xác: đảm bảo sai số không vượt quá ±0,5 giây để tối ưu cho cả người học và các hệ thống tự động

2.1.5. Kết quả kỳ vọng và tiềm năng ứng dụng của bộ dữ liệu phụ đề

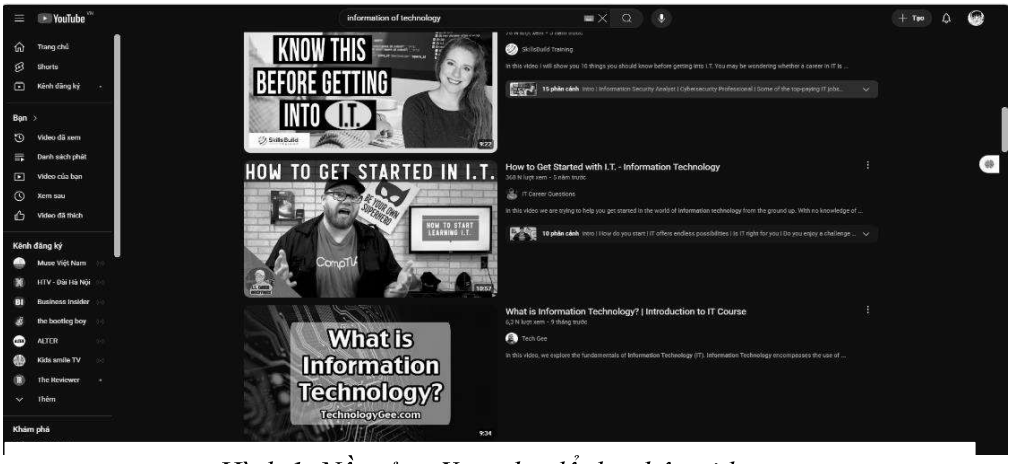
Kết quả từ việc thu thập và xây dựng bộ dữ liệu phụ đề video chuyên ngành CNTT dự kiến sẽ đạt được nhiều thành tựu quan trọng. Nghiên cứu dự kiến tạo ra bộ dữ liệu phụ đề gồm 15k mẫu nội dung giáo dục CNTT đa dạng. Dựa trên bộ dữ liệu này, ít nhất ba mô hình NLP chuyên biệt sẽ được phát triển, bao gồm mô hình tóm tắt, mô hình dịch thuật và mô hình trích xuất thông tin. Kết quả nghiên cứu cũng sẽ góp phần nâng cao khả năng tiếp cận học liệu CNTT cho ba nhóm đối tượng chính: người khiếm thính, người học không bản địa và người có khó khăn học tập.

Ngoài ra, nghiên cứu sẽ cung cấp các công cụ và thuật toán hỗ trợ cho việc xây dựng hệ thống trợ lý học tập thông minh, hệ thống đề xuất học tập cá nhân hóa và công cụ đánh giá tự động. Bộ dữ liệu và các công cụ xử lý liên quan sẽ được công bố để đóng góp cho cộng đồng nghiên cứu, tạo nền tảng cho các nghiên cứu tiếp theo trong lĩnh vực NLP và AI ứng dụng trong giáo dục CNTT. Việc xây dựng bộ dữ liệu phụ đề video chuyên ngành CNTT này không chỉ tạo ra nguồn tài nguyên học tập có giá trị, mà còn mở ra nhiều cơ hội mới cho việc ứng dụng công nghệ xử lý ngôn ngữ tự nhiên trong giáo dục và đào tạo.

2.2. Nguồn thu thập phụ đề video

2.2.1. Phương pháp lựa chọn nền tảng

Sau khi phân tích các nền tảng chia sẻ video phổ biến, YouTube được xác định là nguồn chính để thu thập phụ đề video chuyên ngành công nghệ thông tin (CNTT). Việc lựa chọn này dựa trên một số tiêu chí quan trọng. Trước hết, YouTube cung cấp tính đa dạng nội dung, bao gồm nhiều tài liệu giáo dục về CNTT từ cơ bản đến nâng cao. Khả năng tiếp cận cũng là yếu tố quan trọng khi nền tảng này cung cấp API cho phép trích xuất phụ đề một cách tự động và hiệu quả. Ngoài ra, YouTube hỗ trợ nội dung bằng nhiều ngôn ngữ, trong đó có tiếng Anh và tiếng Việt, đáp ứng đúng yêu cầu của dự án. Độ tin cậy của nguồn cũng được đảm bảo nhờ sự hiện diện của nhiều kênh giáo dục uy tín, với lượng người đăng ký lớn và các đánh giá tích cực từ người xem.



Hình 1. Nền tảng Youtube để thu thập video

2.2.2. Đánh giá và lựa chọn kênh

Độ chính xác của nội dung được đánh giá dựa trên sự so sánh với các nguồn học thuật và ý kiến từ chuyên gia trong lĩnh vực. Chuyên môn của người tạo nội dung được xem xét

thông qua trình độ học vấn, kinh nghiệm thực tiễn và uy tín trong ngành CNTT. Khía cạnh tương tác với cộng đồng được đánh giá bằng phản hồi từ bình luận và mức độ tương tác giữa người tạo nội dung và người học. Cuối cùng, chất lượng sư phạm được đo lường qua khả năng truyền đạt các kiến thức phức tạp một cách dễ hiểu, logic và có hệ thống.

2.2.3. Tiêu chí lựa chọn video chi tiết

Dựa trên mục tiêu xây dựng bộ dữ liệu chất lượng cao, các tiêu chí cụ thể đã được thiết lập để lựa chọn video phù hợp. Về mặt ngôn ngữ, các video được lựa chọn phải có nội dung bằng tiếng Anh hoặc tiếng Việt, với yêu cầu về chất lượng phát âm rõ ràng, tốc độ nói phù hợp và thuật ngữ chuyên ngành được sử dụng chính xác. Các bài giảng chuyên sâu có thể kéo dài hơn nhưng sẽ được phân tách thành các phần nhỏ hơn để đảm bảo hiệu quả tiếp thu.

Phụ đề được trích xuất cần đảm bảo độ chính xác tối thiểu 99% theo tiêu chuẩn của FCC. Phụ đề thủ công được ưu tiên hơn vì thường có độ chính xác cao hơn so với phụ đề tự động. Định dạng phụ đề phải phù hợp với các chuẩn phổ biến như .srt, .vtt hoặc .sbv và phải được đồng bộ chính xác với âm thanh. Nội dung của video phải tập trung nghiêm ngặt vào lĩnh vực CNTT, được xuất bản trong vòng 3 năm trở lại và có cấu trúc rõ ràng bao gồm phần giới thiệu, nội dung chính và kết luận.

2.2.4. Quy trình thu thập và xử lý phụ đề

Quy trình thu thập được thực hiện theo các bước cụ thể. Đầu tiên, các video phù hợp sẽ được xác định dựa trên các tiêu chí đã đề ra. Sau đó, phụ đề sẽ được trích xuất thông qua YouTube API hoặc các công cụ chuyên dụng, sau đó được phân loại theo chủ đề, ngôn ngữ và độ phức tạp. Quá trình xử lý bao gồm việc làm sạch dữ liệu để loại bỏ các thông tin không liên quan như quảng cáo hoặc lỗi chính tả. Định dạng của phụ đề sẽ được chuẩn hóa để đảm bảo tính nhất quán trong bộ dữ liệu. Kiểm tra chất lượng sẽ được thực hiện trên 10% mẫu để đảm bảo độ chính xác, và các thuật ngữ chuyên ngành sẽ được chú thích thêm nếu cần thiết.

Sau khi xử lý, phụ đề sẽ được lưu trữ trong cơ sở dữ liệu có cấu trúc, với các trường thông tin đầy đủ để hỗ trợ tra cứu và phân tích. Dữ liệu sẽ được bảo mật theo các quy định về bản quyền và sử dụng công bằng. Quá trình theo dõi phiên bản cũng sẽ được thực hiện để đảm bảo khả năng kiểm soát và cập nhật kịp thời.

2.2.5. Đánh giá thí điểm

Để đảm bảo chất lượng của quy trình thu thập, nghiên cứu đã tiến hành thí điểm với 100 video từ các kênh đã chọn. Kết quả cho thấy 82% video đáp ứng đầy đủ các tiêu chí đề ra. Tuy nhiên, một số thách thức vẫn tồn tại, đặc biệt là độ chính xác của phụ đề tự động khi xử lý các thuật ngữ kỹ thuật. Phân bố ngôn ngữ trong bộ dữ liệu bao gồm 70% tiếng Anh và 30% tiếng Việt. Về mặt chủ đề, các video được thu thập chủ yếu xoay quanh phát triển web (40%), lập trình (30%), trí tuệ nhân tạo và máy học (15%), còn lại 15% thuộc các lĩnh vực khác. Dựa trên kết quả này, quy trình thu thập đã được điều chỉnh để tăng cường kiểm tra thủ công đối với phụ đề tự động và mở rộng phạm vi tìm kiếm cho một số chủ đề còn thiếu.

2.3. Quá trình thu thập phụ đề video

2.3.1. Quá trình lấy link video và tiêu đề video

Sử dụng API của Google Cloud : API dữ liệu YouTube v3 như Hình 2 và Hình 3 [1]

❖ Các chức năng chính của YouTube Data API v3

Tìm kiếm video: Cho phép tìm kiếm video dựa trên từ khóa, loại video và nhiều tiêu chí khác.

Truy xuất thông tin video: Truy xuất thông tin chi tiết về video như tiêu đề, mô tả, kênh, số lượt xem, ngày đăng tải và các thông tin khác.

Truy xuất thông tin kênh: Lấy thông tin về kênh YouTube như tên kênh, số lượng người theo dõi, danh sách các video hoặc danh sách phát của kênh.

Lấy danh sách phát: Truy xuất các video trong một playlist và các thông tin liên quan.

Quản lý nội dung cá nhân: API cũng cho phép người dùng đã xác thực thêm, chỉnh sửa hoặc xóa nội dung (video, playlist, bình luận) trên tài khoản YouTube của họ.

```
from googleapiclient.discovery import build
import pandas as pd
import os

# Khởi tạo YouTube API
api_key = 'AIzaSyDyz6BBB11Kz5X0WKF4xYnZKmQycE1qXPA'
youtube = build('youtube', 'v3', developerKey=api_key)
```

Hình 2. Sử dụng Google API Client để tạo một đối tượng youtube có thể tương tác với YouTube Data API.

```
# Hàm kiểm tra nếu video liên quan đến IT, Code, AI...
def is_relevant(video):
    keywords = ["IT", "Information Technology", "AI", "Artificial Intelligence", "Programming", "Coding", "Software Development", "Machine Learning",
    title = video['snippet']['title'].lower()
```

Hình 3. Kiểm tra tiêu đề của mỗi video xem có chứa các từ khóa liên quan đến IT, lập trình, AI, và các lĩnh vực công nghệ khác hay không.

Sử dụng yt-dlp [2]

❖ Các tính năng chính của yt-dlp:

Tải về video từ nhiều trang web: yt-dlp hỗ trợ nhiều trang web, không chỉ YouTube mà còn các trang khác như Vimeo, Dailymotion, Facebook, Twitter, SoundCloud và nhiều trang video, nhạc khác.

Tải nhiều định dạng: Công cụ này hỗ trợ tải nhiều định dạng video (MP4, MKV, AVI, v.v.) hoặc chỉ tải audio (MP3, M4A). Bạn có thể chỉ định định dạng bạn muốn hoặc để yt-dlp chọn định dạng tốt nhất.

Tải danh sách phát (playlist): yt-dlp có khả năng tải toàn bộ danh sách phát (playlist) từ các trang như YouTube. Bạn có thể tải tất cả các video hoặc chỉ tải một phần danh sách.

Tải phụ đề: yt-dlp có thể tải cả phụ đề nếu chúng có sẵn trên video, giúp bạn thu thập thông tin văn bản từ video.

Tải metadata: Nó hỗ trợ tải metadata của video như tiêu đề, mô tả, thời gian đăng tải, và thậm chí cả ảnh thumbnail của video.

Tùy chỉnh chất lượng tải về: Bạn có thể chỉ định chất lượng video muốn tải (720p, 1080p, 4K), hoặc để yt-dlp tự động chọn chất lượng tốt nhất có sẵn.

Bypass giới hạn địa lý: yt-dlp có thể sử dụng các phương thức như proxy để vượt qua các hạn chế về địa lý khi tải nội dung từ những trang có kiểm soát khu vực.

Tải về nhiều video cùng lúc: Bạn có thể sử dụng yt-dlp để tải hàng loạt video một cách đồng thời, giúp tiết kiệm thời gian khi thu thập dữ liệu lớn.

Bổ sung tính năng mới: yt-dlp đã cải tiến nhiều tính năng so với youtube-dl như hỗ trợ tải về từ các dịch vụ phát trực tuyến (streaming) cao cấp và những tùy chọn như merge video và audio vào cùng một file, speed throttling, và các cách thức tối ưu hiệu suất tải về.

Sử dụng các lệnh hệ thống: yt-dlp được thực thi qua lệnh dòng lệnh, dễ dàng tích hợp với các ngôn ngữ lập trình (như Python) để tự động hóa quá trình tải video và xử lý dữ liệu từ nhiều video cùng lúc.

```
# Lệnh yt-dlp để lấy danh sách video
command = [
    "yt-dlp",
    "--flat-playlist",
    "--dump-json",
    channel_url
]
```

Hình 4. Tạo một danh sách các tham số lệnh để gọi yt-dlp

```
# Đường dẫn đến kênh YouTube
channel_url = "https://www.youtube.com/@TwoMinutePapers"
```

Hình 5 . Đường dẫn đến kênh YouTube mà bạn muốn lấy danh sách video

2.3.2. Quá trình lấy phụ đề video

Sử dụng youtube-transcript-api [3]

youtube-transcript-api là một thư viện Python được phát triển để giúp bạn dễ dàng lấy phụ đề từ các video YouTube. Thư viện này hoạt động bằng cách sử dụng API không chính thức của YouTube, giúp bạn truy xuất phụ đề của video một cách đơn giản mà không cần thiết phải thông qua các quy trình phức tạp như lấy API key từ YouTube.

Các điểm nổi bật của thư viện:

- Không cần API key từ YouTube: Một trong những ưu điểm chính của youtube-transcript-api là bạn không cần phải lấy API key từ YouTube Developer Console. Điều này giúp việc truy cập phụ đề trở nên thuận tiện hơn.
- Hỗ trợ nhiều ngôn ngữ: Bạn có thể yêu cầu phụ đề bằng nhiều ngôn ngữ khác nhau. Thư viện sẽ kiểm tra xem ngôn ngữ được yêu cầu có sẵn không và trả về phụ đề tương ứng. Nếu phụ đề không có trong ngôn ngữ đầu tiên, bạn có thể chỉ định một danh sách ngôn ngữ ưu tiên để kiểm tra.
- Cách thức hoạt động:Thư viện sử dụng ID của video YouTube để lấy phụ đề và trả về dưới dạng danh sách các đoạn văn bản kèm thời gian xuất hiện trong video.

```
# Hàm lấy phụ đề từ video
def get_subtitles(video_id, languages=['vi', 'en']):
    for lang in languages:
        try:
            transcript = YouTubeTranscriptApi.get_transcript(video_id, languages=[lang])
            subtitles = " ".join([t['text'] for t in transcript])
            return subtitles
        except Exception as e:
            print(f"Không thể lấy phụ đề bằng ngôn ngữ {lang} cho video {video_id}: {e}")
    print(f"Không có phụ đề nào có sẵn cho video {video_id}.")
    return None
```

Hình 6. Lấy phụ đề của những video (đã thu thập) với 2 loại ngôn ngữ Tiếng Việt và Tiếng Anh

3. KẾT LUẬN

Nghiên cứu này đã phát triển một phương pháp tự động thu thập và xử lý phụ đề từ các video chuyên ngành Công nghệ Thông tin (CNTT) trên YouTube, nhằm tạo ra bộ dữ liệu chất lượng cao phục vụ cho các nhiệm vụ tóm tắt tự động. Việc lựa chọn YouTube làm nguồn thu thập chính được hỗ trợ bởi tính đa dạng và khả năng tiếp cận rộng rãi của nền tảng này, đáp ứng nhiều phong cách học tập khác nhau, bao gồm thị giác, thính giác và vận động [4].

Quy trình thu thập và xử lý phụ đề được thiết kế theo hướng tự động hóa, từ việc trích xuất dữ liệu thông qua YouTube Data API, làm sạch và chuẩn hóa phụ đề, đến kiểm tra chất lượng và lưu trữ trong cơ sở dữ liệu có cấu trúc. Kết quả thí điểm trên 100 video cho thấy tỷ lệ phù hợp đạt 82%, với sự phân bố hợp lý giữa các chủ đề như phát triển web, lập trình và AI/ML.

Bộ dữ liệu thu được không chỉ hữu ích cho việc phát triển các mô hình tóm tắt tự động mà còn có giá trị trong việc huấn luyện và đánh giá các hệ thống xử lý ngôn ngữ tự nhiên. Các nghiên cứu đã chỉ ra rằng việc sử dụng phụ đề trong video giáo dục có thể cải thiện đáng kể khả năng hiểu bài và giảm tải nhận thức cho người học [5].

Trong tương lai, nghiên cứu sẽ tiếp tục mở rộng quy mô bộ dữ liệu và cải tiến mô hình tóm tắt để nâng cao chất lượng đầu ra, đặc biệt trong việc xử lý các thuật ngữ chuyên ngành và duy trì tính nhất quán ngôn ngữ. Điều này phù hợp với xu hướng hiện nay trong việc sử dụng các công cụ tự động để trích xuất và xử lý phụ đề từ video, nhằm hỗ trợ hiệu quả cho các ứng dụng giáo dục và nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Google, “Let users watch, find, and manage YouTube content,” [Trực tuyến], Available: https://developers.google.com/youtube/.
2. Subrat-lima, “yt-dlp 2025.3.31,” [Trực tuyến]. Available: https://pypi.org/project/yt-dlp/.
3. “youtube-transcript-api 1.0.3,” [Trực tuyến], Available: https://pypi.org/project/youtube-transcript-api/.
4. University, “YouTube and Education: Leveraging the Platform for Learning Beyond the Classroom,” [Trực tuyến]. Available: https://www.lordsuni.edu.in/blog/youtube-and-education.
5. PMC, “The effects of using an auto-subtitle system in educational videos to facilitate learning for secondary school students: learning comprehension, cognitive load, and satisfaction,” [Trực tuyến], Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9831372/.
6. D. Rosa, “How IDC’s Industry CloudPath & SaaSPath Surveys Can Inform Your Cloud/SaaS Strategy,” 4 9 2019. [Trực tuyến]. Available: https://blogs.idc.com/2019/09/04/how-idcs-industry-cloudpath-saaspath-surveys-can-inform-your-cloud-saas-strategy/.
7. Kimbrough, “AI is shifting the workplace skillset. But human skills still count,” 21 1 2025. [Trực tuyến]. Available: https://www.weforum.org/stories/2025/01/ai-workplace-skills/.
8. Muscanell, “EDUCAUSE QuickPoll Results: Transforming Teaching and Learning with a Digital Learning Strategy,” 22 8 2022. [Trực tuyến]. Available: https://er.educause.edu/articles/2022/8/educause-quickpoll-results-transforming-teaching-and-learning-with-a-digital-learning-strategy.

TOWARDS THE DEVELOPMENT OF A SUBTITLE DATASET FOR EDUCATIONAL VIDEOS IN INFORMATION TECHNOLOGY

Abstract: *This study aims to construct a domain-specific dataset of video subtitles in the field of Information Technology (IT) to enhance access to educational resources and support the development of natural language processing (NLP) applications in education. A systematic methodology is proposed for data collection and processing, encompassing source selection, subtitle extraction, data cleaning, normalization, and quality assurance. The resulting dataset possesses strong academic value and is intended to serve as a foundational resource for further research and practical applications in IT education.*

Keywords: *educational subtitles; natural language processing; artificial intelligence in education; digital learning resources in IT.*

.