HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN NGỌC HÙNG ANH

NGHIÊN CỨU GIẢI PHÁP PHÂN TÍCH HÀNH VI NGƯỜI DÙNG QUA MẠNG HỌC SÂU NHẰM THIẾT KẾ GIẢI THUẬT TƯ VẤN KÊNH CHO NGƯỜI XEM TRUYỀN HÌNH

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH – NĂM 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Ngọc Hùng Anh

NGHIÊN CỨU GIẢI PHÁP PHÂN TÍCH HÀNH VI NGƯỜI DÙNG QUA MẠNG HỌC SÂU NHẰM THIẾT KẾ GIẢI THUẬT TƯ VẤN KÊNH CHO NGƯỜI XEM TRUYỀN HÌNH

Chuyên ngành : HỆ THỐNG THÔNG TIN

Mã số: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TS. TRẦN THU HÀ

TP. HỒ CHÍ MINH – NĂM 2022

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của bản thân tôi. Các số liệu, kết quả được trình bày trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào trước đây.

Tp.HCM, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Ngọc Hùng Anh

LỜI CẢM ƠN

Em xin chân thành cảm ơn **PGS.TS Trần Thu Hà**, Khoa điện điện tử, Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh đã tận tình chỉ dạy và hướng dẫn cho em trong việc lựa chọn đề tài, thực hiện đề tài và viết báo cáo luận văn, giúp cho em có thể hoàn thành tốt luận văn này.

Em xin chân thành cảm ơn các Thầy Cô trong Khoa Công nghệ thông tin là những người giảng dạy em, đặc biệt là các Thầy Cô trong Khoa Sau đại học đã tận tình dạy dỗ và chỉ bảo em trong suốt 2 năm học.

Cuối cùng em xin cảm ơn gia đình, bạn bè, những người đã luôn bên cạnh động viên em những lúc khó khăn và giúp đỡ em trong suốt thời gian học tập và nghiên cứu, tạo mọi điều kiện tốt nhất để cho em có thể hoàn thành tốt luận văn của mình.

Mặc dù đã cố gắng hoàn thành nghiên cứu trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự thông cảm của quí Thầy Cô và các bạn.

Em xin chân thành cảm ơn !

Tp.HCM, ngày 25 tháng 01 năm 2022

Học viên thực hiện luận văn

Nguyễn Ngọc Hùng Anh

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
OTT	Over The Top	Truyền hình số qua mạng Internet
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformer
IPTV	Internet Protocol TV	Truyền hình Internet
LSTM	Long-Short Term Memory	Mạng bộ nhớ dài-ngắn
BiLSTM	Bidirectional long short-term memory	Mạng bộ nhớ dài-ngắn hai chiều
SRM	Structural rRisk rMinimization	Cực tiểu hóa rủi ro có cấu trúc
SVM	Support Vector machine	Máy vector hỗ trợ
VoD	Video on Demand	Video theo yêu cầu
NSP	Next Sentence Prediction	Dự đoán câu tiếp theo
MLM	Masked Language Modeling	Tạo mô hình ngôn ngữ có mặt nạ
STB	Set-top-box	Đầu thu tín hiệu
PC	Personal Computer	Máy tính cá nhân
CND	Content Delivery Network	Mạng lưới trung chuyển phân phối nội dung
CMS	Content Management System	Hệ thống quản lý nội dung
IP	Internet Protocol	Các giao thức truyền tải thông tin trên Internet

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, Ngành Công nghệ thông tin đã và đang được phát triển rất mạnh về phần cứng và cũng như phần mềm. Với sự phát triển đó, có một lĩnh vực cũng đang phát triển rất mạnh, cũng là xu thế trong tương lai và là một sự kết hợp giữa sự phát triển của phần cứng lẫn phần mềm đó là lĩnh vực dịch vụ phát sóng Truyền hình trên Internet.

Để duy trì dịch vụ Truyền hình trên Internet, mô hình OTT (Over The Top) là giải pháp cung cấp nội dung cho người sử dụng dựa trên nền tảng Internet cung cấp bởi bên thứ ba. Công nghệ OTT cho phép cung cấp các nguồn Truyền hình có nội dung phong phú đa dạng theo yêu cầu của người sử dụng vào bất kì thời điểm nào, tại bấy kì nơi đâu chỉ với một thiết bị phù hợp với ứng dụng và có kết nối Internet. [1]

Trên thế giới, công nghệ OTT đã làm thay đổi bộ mặt của dịch vụ truyền hình số cổ điển. Cùng với sự phát triển của các thiết bị công nghệ hiện đại như điện thoại, máy tính, Smart TV và các phương tiện kỹ thuật số.

Nhằm giúp cho người sử dụng có thể nhanh chóng tìm ra kênh / nội dung muốn xem, mô hình OTT đã có những tiện ích như sau:

- Tạo ứng dụng chương trình xem lại kênh vừa mới xem ngay trước đó. Tâm lý là người xem thường chọn cho mình thêm một chương trình dự bị khi kênh đang xem không còn cuốn hút (do quảng cáo, do trục trặc kỹ thuật), chính vì thế việc luân chuyển giữa hai kênh thường xem, chỉ sử dụng một nút nhấn là cách rất hiệu quả giúp người xem nhanh chóng xem được chọn lựa của mình.
- Tạo danh sách các kênh yêu thích, giảm số lượng hàng trăm kênh xuống thành một vài kênh mà người xem quan tâm nhất.
- Tạo các chủ đề để phân loại các chương trình xem lại như kênh tổng hợp, ca nhạc, phim, v.v… Nhờ đó mà người xem sẽ nhanh chóng hơn khi chọn được chủ đề và chương trình để xem.

Tiện ích thứ 3 chỉ áp dụng được cho các nội dung xem lại, VoD (Video on Demand). Đối với các kênh truyền hình trực tiếp, chưa thể xem các chương trình phát sóng theo chủ đề riêng. Việc sử dụng lịch phát sóng truyền thống vẫn là giải pháp được áp dụng rộng rãi ở các kênh truyền hình: các chương trình phát sóng được liệt kê theo lần lượt theo thứ tự thời gian và cho từng đài / kênh phát sóng. Người sử dụng phải chọn kênh phát sóng để xem chương trình đang phát có đúng chủ đề mình cần xem hay không. Thông tin về nội dung chương trình phát sóng có thể được mô tả trong lịch phát sóng. Tuy nhiên người xem phải đọc một cách "thủ công" tất cả thông tin này cho từng chương trình phát sóng để tìm ra đúng nội dung yêu thích.

Với hạn chế nêu trên khi tìm kiếm chương trình truyền hình muốn xem, chúng ta có thể ứng dụng những tiến bộ của công nghệ để cung cấp dịch vụ cho người dùng một cách tối ưu hơn nên em chọn đề tài **"Nghiên cứu giải pháp phân tích hành vi người dùng qua mạng học sâu nhằm thiết kế giải thuật tư vấn kênh cho người xem truyền hình"** cho luận văn Thạc sĩ này. Mục đích là cải thiện chất lượng thời gian tìm kiếm thông tin của chủ đề và gợi ý những nội dung tiếp theo giúp cho người xem dễ dàng xem những chủ đề yêu thích một cách nhanh nhất.

2. Mục đích nghiên cứu

Nghiên cứu phân tích hành vi người dùng qua mạng học sâu và thiết kế giải thuật tư vấn kênh cho người xem truyền hình:

- Nghiên cứu, phân loại đoạn văn tiếp nhận đầu vào và dùng các mô hình phân tích biết trước để xử lý đoạn văn của chương trình truyền hình và phân loại nhóm theo tựa đề của chương trình phát sóng trong lịch phát sóng truyền thống và gán thành các nhãn là tên của chủ đề trong giao diện dịch vụ tìm kiếm. Đây là một giải pháp nâng cao chất lượng dịch vụ trong Truyền hình sẽ tiết kiệm thời gian tra cứu kênh và nội dung theo chủ đề cho người xem.
- Nghiên cứu ứng dụng thuật toán Kmeans trên cơ sở các quy luật xác định, đề xuất các tiêu chí để đánh giá, phân loại nội dung, tần suất xuất hiện của

các cụm từ, các cấu trúc văn phạm, cách dùng từ, các diễn giải để làm cơ sở xác định chủ đề của nội dung Truyền hình. [4]

- Nghiên cứu và thiết kế giải thuật phân biệt câu từ, ngữ pháp, động từ , danh từ thuộc cấu trúc câu và tiến hành "đào tạo" các thuộc tính. Các nội dung sẽ được huấn luyện và gán vào một chủ đề tương ứng [2].
- Tiến hành thử nghiệm sản phẩm giúp người dùng có thể tìm kiếm được kênh truyền hình và biết thông tin kênh sẽ có nội dung mong muốn xem tiết kiệm thời gian tạo cảm giác thoải mái cho người dung đầu cuối khi giải trí.

3. Đối tượng và phạm vi nghiên cứu

❖ ***Đối tượng nghiên cứu:***

Biến đổi dữ liệu thô thu được từ các trang web có lịch phát sóng Truyền hình để phục vụ mục đích nghiên cứu. [3]

Sử dụng thuật toán K-means clustering để phân loại và bổ sung theo luật xác định để tìm ra chủ đề của chương trình Truyền hình.

Sử dụng phương pháp tự động phân loại và bổ sung theo từng chủ đề của chương trình Truyền hình dựa vào mô hình máy học PhoBERT. [4]

So sánh các phương pháp phân loại đoạn văn như: SVM, Bert, PhoBERT.

❖ **Phạm vi nghiên cứu:**

Dựa vào các quy luật xác định để phân tích được số lần xuất hiện của các cụm từ, cấu trúc văn phạm của người dùng yêu cầu để làm cơ sở xác định cho việc quyết định nhóm gợi ý cho người xem.

Dựa vào hỗ trợ của mô hình máy học PhoBERT để phân tích tự động nội dung chủ đề và bổ sung theo từng chủ đề yêu thích của người xem.

Mô hình OTT được chia thành ba thành phần chính, thực hiện những chức năng một cách tuần tự như sau:

- Thu thập thông tin từ trạng thái của hệ thống.
- Nhận yêu cầu từ bộ phận người dùng, xây dựng mô hình và ra quyết định.
- Nhận lệnh và thực thi.

4. Phương pháp nghiên cứu

Luận văn này sử dụng các phương pháp nghiên cứu lý thuyết và kết hợp với xây dựng ứng dụng thử nghiệm:

- Thu thập các tài liệu, thông tin có liên quan tới đề tài để phục vụ nghiên cứu.
- Ứng dụng các công nghệ lập trình python và các công nghệ trong lĩnh vực máy học như: BERT, PhoBERT, v.v… để so sánh, phát triển hệ thống thử nghiệm
- Tiến hành đánh giá kết quả thử nghiệm, đưa ra hướng phát triển mở rộng của đề tài để đáp ứng những nhu cầu triển khai thực tế.

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

Chương này luận văn giới thiệu khái quát về vai trò của OTT trong dịch vụ truyền hình Internet. Hiệu quả của tính năng trong quá trình điều chỉnh nội dung để thích ứng với nguồn phát. Phân loại nội dung của chương trình phát theo từng nhóm của chủ đề. Hiệu quả của việc phân loại chương trình theo nội dung truyền tải. Giúp cho chúng ta thấy được tầm quan trọng của việc phân loại nội dụng của kênh Truyền hình. Gợi ý cho người xem thông qua sở thích và thói quen của họ.

1.1 Tổng quan về mô hình OTT

Các dịch vụ ứng dụng đa phương tiện miễn phí trên các thiết di động đã thu hút hàng triệu người Việt Nam, đặt biệt là các giới trẻ. Các dịch vụ này đã làm cho các nhà mạng trong nước lo lắng về sự cạnh tranh, chia sẻ các doanh thu. Tuy nhiên với sự phát triển mạnh mẽ của các dịch vụ truyền hình Internet đang là xu hướng trong tương lai và không thể tránh khỏi sự cạnh tranh hoặc hợp tác giữa các nhà mạng trong nước.

Dịch vụ truyền hình Internet là một trong những dịch vụ đã thay đổi rất nhiều dựa vào sự thay đổi về thói quen và hành vi tiếp cận của người dùng. Đặt biệt với nhu cầu Internet đang phát triển rất mạnh, người dùng luôn lựa chọn những dịch vụ dựa theo sở thích cá nhân trên thiết bị TV thông minh hoặc điện thoại thông minh.

Nhờ sự phát triển Internet làm cho dịch vụ truyền hình trở nên phổ biến và ngày càng gần hơn với người dùng. Các chương trình truyền hình ngày nay luôn phát trực tuyến trên các thiết bị thông minh giúp cho người xem có thể xem và lựa chọn những chương trình yêu thích của họ mọi lúc mọi nơi.

Ứng dụng OTT (Over The Top) là giải pháp cung cấp các nội dung cho người dùng như âm thanh, hình ảnh trên nền tảng Internet độc lập, với mô hình công nghệ OTT, những nội dung truyền hình được phân phối qua nhiều hạ tầng Internet, không nhất thiết sở hữu bởi nhà cung cấp dịch vụ. Đây là điểm khác biệt so với các dịch vụ truyền thống như truyền hình cáp, truyền hình vệ tinh. [5]

Với sự phát triển của các thiết bị công nghệ như smartphone, Smart TV đã làm thay đổi các nhà mạng cũng như dịch vụ truyền hình, đặt biệt là trong khoảng 10 năm qua, và chắc chắn sẽ còn rất nhiều thay đổi trong những năm tiếp theo. Từ đó mô hình OTT đang ngày càng sử dụng phổ biến trong lĩnh vực Internet và đã mở ra nhiều cơ hội mới cho các nhà cung cấp dịch vụ truyền hình như Netflix, VTVGo, SCTV Online, v.v… [6]

Tại Việt Nam dịch vụ truyền hình Internet phát qua Smart TV và ứng dụng truyền hình phát trên các thiết bị di động ngày càng phổ biến và tăng mạnh, các nhà cung cấp truyền hình OTT luôn đầu tư và phát triển với nội dung chất lượng cao và đa dạng hơn, giúp cho người dùng dễ dàng xem và chọn lựa nội dung mình yêu thích dễ dàng nhất.

1.2 Mô hình IPTV truyền thống

1.2.1 Sơ lược về IPTV

Sự phát triển mạnh mẽ của mạng Internet toàn cầu đã góp phần khai sinh ra một hình thức truyền hình hoàn toàn mới và đầy hứa hẹn. Đó là truyền hình Internet "Internet Protocol Television" (IPTV). Mặc dù ra đời từ cách đây hơn một thập kỷ nhưng có thể thấy IPTV hầu như không thể phát triển mạnh mẽ như mong đợi bởi trong quá khứ do điều kiện hạ tầng và băng thông mạng chưa cho phép loại hình truyền hình mới này phát huy hết lợi thế. Chính vì thế mà IPTV vẫn còn nhường bước so với truyền hình truyền thống và truyền hình cáp.

Trong những năm gần đây mạng Internet đã có những bước phát triển vượt bậc. Trong đó đáng chú ý nhất là sự phổ biến của mạng băng rộng với tốc độ kết nối ngày càng nhanh hơn. Ở một số quốc gia như Hàn Quốc cáp quang đã được kéo đến tận từng nhà. Đây là nền tảng giúp IPTV bắt đầu có bước phát triển mạnh mẽ.

IPTV có thể xem là thế hệ tiền thân của truyền hình trên nền tảng OTT. Trên hệ thống IPTV, dịch vụ truyền hình số được được cung cấp qua thiết bị đầu cuối Set-top-box (STB). Qua thiết bị này, thuê bao có thể xem các kênh, thực hiện dịch vụ thuê bao cũng như các dịch vụ tương tác đa phương tiện khác thông qua nền tảng kết nối trực tiếp – quản lý bởi chính nhà cung cấp dịch vụ (managed IP). Bản chất kết

nối giữa STB và nhà cung cấp dịch vụ là dựa trên nền tảng IP, nên dịch vụ IPTV có thể dễ dàng được cung cấp cùng với dịch vụ Internet khác như truy cập trang Web, điện thoại qua Internet, v.v… [7]

- Hỗ trợ truyền hình có tính tương tác 2 chiều: tạo điền kiện cho việc cung cấp đa dạng các ứng dụng truyền hình có tính tương tác cao như truyền hình trực tiếp với nhiều góc quay, truyền hình có độ nét cao theo yêu cầu, các trò chơi truyền hình tương tác, v.v…
- Xem lại chương trình của kênh truyền hình: kết hợp với chức năng ghi hình cho phép người dùng xem lại chương trình đã phát sóng ở một thời điểm khác trước đây.
- Cải thiện trải nghiệm riêng biệt khi xem truyền hình: nhờ tương tác 2 chiều với nhà cung cấp dịch vụ thông qua STB, người dùng có thể chọn lựa kênh muốn xem và thời gian xem cho phù hợp với thị hiếu của mình.
- Sử dụng băng thông một cách hiệu quả: công nghệ IPTV bảo đảm chỉ phát kênh lên hạ tầng truyền dẫn khi có người yêu cầu. Chính thế dù có khả năng cung cấp rất nhiều chương trình cùng một thời điểm, băng thông của hạ tầng cũng được sử dụng một cách hợp lý.
- Giải trí thư giãn xem truyền hình qua nhiều thiết bị đầu cuối, hệ thống IPTV cung cấp nội dung không chỉ trên TV mà còn có thể trên PC hay trên điện thoại thông minh kết nối trực tiếp với mạng nội bộ của STB.

1.2.2 Kiến trúc cơ bản của hệ thống IPTV

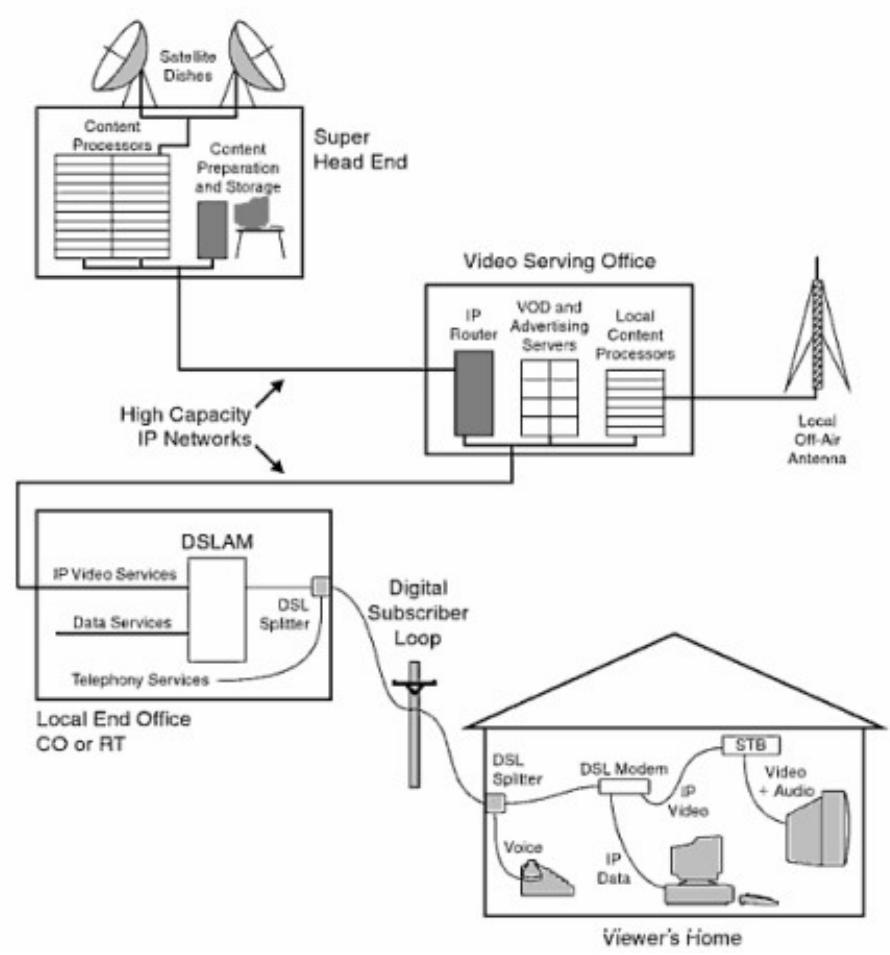
Super head-end (đầu nạp tải trung tâm): nơi tập trung các kênh mà dịch vụ IPTV muốn cung cấp cho khách hàng.

Core network (mạng lưới trung tâm): tốc độ mạng cao, truyền tải các kênh đến các đầu nạp tải khu vực.

Access network (mạng lưới thuê bao) bảo đảm kết nối từ đầu nạp khu vực đến người dùng đầu cuối.

Regional head-end (đầu nạp khu vực): có thêm các kênh khu vực được đưa vào gói kênh phát cho người dùng đầu cuối.

Customer premises (thiết bị người dùng đầu cuối): là hệ thống mạng nội bộ cài ở người dùng, nơi luồng IPTV kết thúc và được trình chiếu. [8]



Hình 1.1: Các thành phần cơ bản của hệ thống IPTV

1.2.3 Sự phát triển của IPTV trong giai đoạn hiện tại

Mặc dù IPTV tồn tại trên nền tảng Internet, nhưng không có nghĩa là dịch vụ này sẽ khả thi với mọi chất lượng Intenet. Chỉ khi Internet với băng thông rộng được phổ cập, IPTV mới được đưa vào ứng dụng rộng rãi vì bản chất tiêu thụ nhiều băng thông của tín hiệu hình ảnh. Hiện tại có 2 dạng để xem IPTV: qua STB hay qua PC được trang bị ứng dụng phù hợp. Nhiều nhà cung cấp IPTV cũng phục vụ luôn dịch vụ điện thoại và truy cập Internet, tạo nên gói dịch vụ đồng thời có 3 tiện ích (tripple play) trên hạ tầng mạng tốc độ cao. [9]

Nhu cầu tăng vọt của Internet trong đời sống hàng ngày và cả trong công việc đã làm cho hạ tầng Internet phát triển nhanh cả về phạm vi phủ mạng lưới, cả về chất lượng mạng. Tận dụng được nền tảng này, IPTV đã dễ dàng hơn trong việc phát triển thị trường mà không cần thêm đầu tư quan trọng cho các hạ tầng chuyên biệt chỉ cho truyền hình. Đấy là lý do mà giai đoạn trước 2010, đánh giá là thời kỳ hoàng kim của IPTV. Giai đoạn này được coi là làn sóng thứ 2 trong ngành công nghiệp truyền hình. Làn sóng thứ nhất là giai đoạn chuyển đổi từ đồng dạng sang số hóa của truyền hình đại chúng.

1.3 Các khó khăn thách thức trong dịch vụ truyền hình Internet

Các nhà cung cấp truyền hình lớn như VTV, VTC, K+, SCTC hoặc các doanh nghiệp trong và ngoài nước như FPT, VNPT, iFlix, Netflix điều tham gia vào cuộc canh tranh cung cấp các gói sản phẩm truyền hình OTT nhằm để đáp ứng được nhu cầu cần thiết của người tiêu dùng.

Trong thực tế, những thách thức lớn cho các nhà mạng cung cấp dịch vụ truyền hình OTT hiện nay đó chính sự thay đổi thói quen hành vi của người dùng và sự phát triển của thiết bị công nghệ.

Truyền hình OTT là lĩnh vực được ứng dụng nhiều nhất bởi việc cung cấp các nội dung truyền hình trực tuyến và các Video. Ưu thế lớn nhất của công nghệ OTT là việc cho phép cung cấp các nguồn nội dung phong phú và đa dạng theo nhu cầu của người dùng. Trong cuộc sống hiện đại ngày nay, người dùng thường thích được xem Truyền hình mọi lúc mọi nơi, theo mong muốn và sở thích, chứ không muốn phụ thuộc vào khung giờ cố định như xem truyền hình như trên TV truyền thốn. [10]

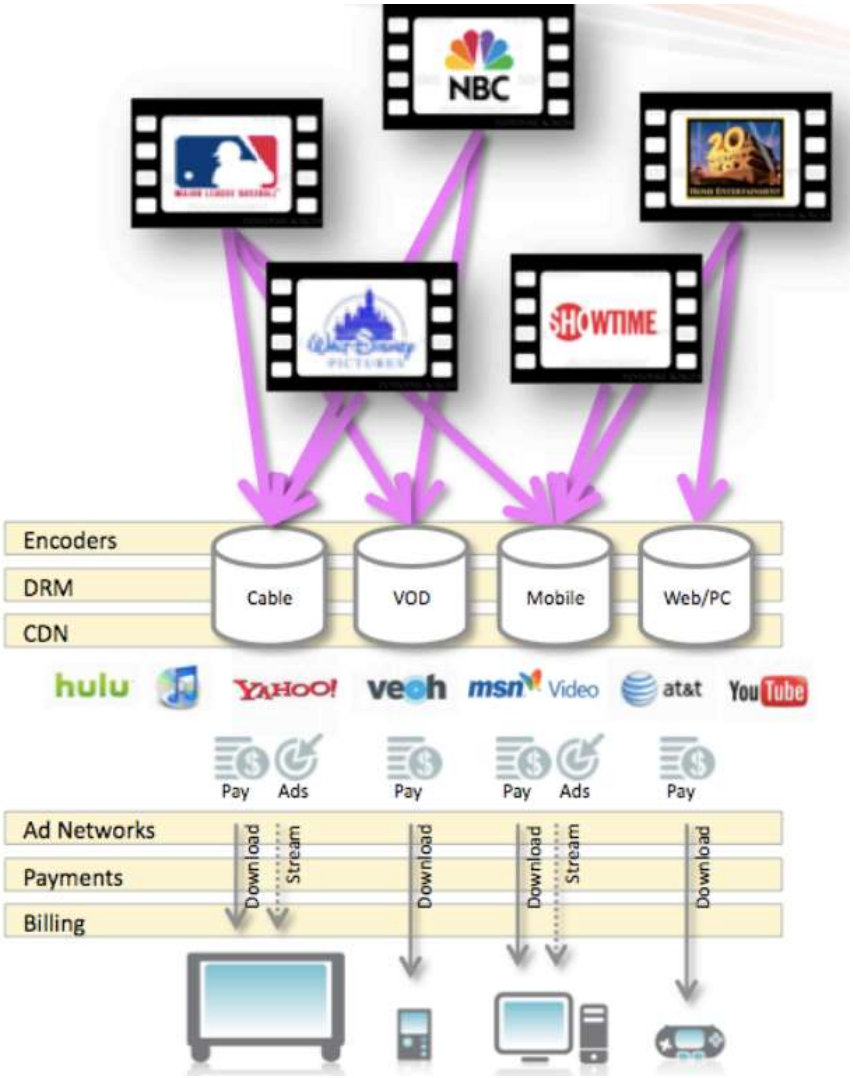
Với những tính năng trên, cùng với những nhu cầu cao của người dùng, mô hình OTT luôn nâng cao chất lượng hình ảnh truyền dẫn, thời gian lựa chọn thay đổi nội dung với băng thông có độ trễ nhỏ hơn 10 giây. Hình ảnh không bị đứng khi thiết bị đầu cuối bị giảm đột ngột. Giao diện hiển thị danh sách kênh được nhóm lại thành các nội dung yêu thích của người dùng, để tiết kiệm thời gian. Kênh đang trình chiếu sẽ nhóm lại thành chủ đề đang chiếu và gợi ý cho người dùng những nội dung tương tự tiếp theo.

Các bước kỹ thuật cũng như dịch vụ kinh doanh chính của một mô hình OTT tiêu biểu. Với bất cứ mô hình nào, các đặc điểm chính của việc triển khai OTT luôn đòi hỏi giải pháp cho các vấn đề sau:

- Số lượng truy cập lớn: không quá bất thường là hiện tượng các gói OTT tạo ra hơn 2,5 triệu người xem trong những tuần đầu triển khai.
- Mô hình mua bản quyền xem truyền hình: có thể mua bản quyền xem phim trên truyền hình tại 1 thiết bị và xem phim đấy qua các thiết bị khác trong nhà.
- Mô hinh OTT theo cơ chế bảo mật, chỉ cho phép người dùng đã có bản quyền xem có thể tận hưởng các phim có trong chương trình TV.

Trong quá trình khảo sát chi tiết các môi trường phát triển OTT khác nhau, và đã diễn giải các vấn đề trên thành các thách thức như sau:

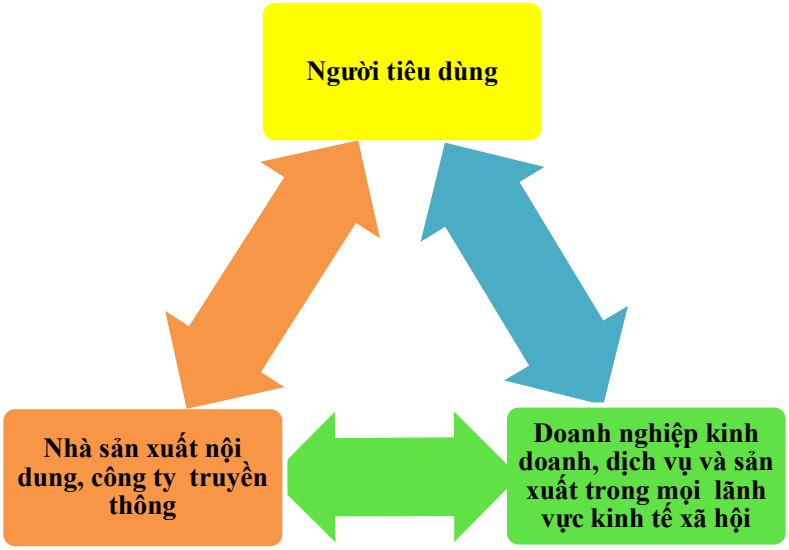
- Khả năng cung cấp nội dung từ nhiều nguồn khác nhau và cho nhiều định dạng cũng như độ phân giải khác nhau.
- Sự đa dạng về số lượng, chất lượng và sự hỗ trợ tính năng khác nhau của thiết bị đầu cuối.
- Tính năng bảo mật nội dung, sự linh động trong việc mua quyền sử dụng.
- Khả năng tích hợp với các hệ thống hỗ trợ vệ tinh đang hoạt động với dịch vụ IPTV như CDN, CMS.
- Khả năng tìm kiếm, phát hiện và nhận tư vấn để có thể tìm ra các nội dung phù hợp.



Hình 1.2: Các giai đoạn chính của một dịch vụ OTT

Khác với mô hình truyền hình đại chúng, chỉ có chi phí cố định không phụ thuộc số lượng người xem, nhà cung cấp OTT phải chú ý sự tăng trưởng của chi phí theo tổng số người sử dụng. Việc tăng số người sử dụng vẫn phải được ưu tiên và điều này phụ thuộc rất nhiều vào trải nghiệm dịch vụ, tiện ích dịch vụ cung cấp cho người dùng đầu cuối. Sự gia tăng các dịch vụ đính kèm trong truyền hình sẽ thu hút nguồn quảng cáo khổng lồ khi số người kết nối tăng khả năng quảng bá rộng mở sẽ càng thu hút người xem truyền hình và thu hút luôn các doanh nghiệp có nhu cầu quảng bá sản phẩm của mình, thúc đẩy mọi mặt kinh doanh, dịch vụ sản xuất tiêu dùng cho xã hội. Mối liên kết tương quan cộng sinh kết hợp chặt chẽ có tương tác

hai chiều, người tiêu dùng sẽ được sử dụng dịch vụ tốt nhất để nhận thông tin và giải trí, nhà sản xuất phim, sản xuất nội dung, các doanh nghiệp truyền thông sẽ cung cấp dịch vụ đáp ứng nhu cầu cho người sử dụng và hợp tác cùng các doanh nghiệp sản xuất kinh doanh dịch vụ của các ngành nghề khác thông qua quảng bá truyền thông với mục tiêu đáp ứng nhu cầu của người tiêu dùng giúp thúc đẩy toàn diện nền kinh tế xã hội. [11]



Hình 1.3: Mối liên kết tương quan giữa người tiêu dùng và doanh nghiệp

1.4 Các phương pháp phân loại văn bản

Bài toán mô hình phân loại văn bản thường có hai cách phân loại khác nhau là: phân loại dựa trên luật và phân loại dựa trên máy học.

Phân loại dựa trên luật là cách phân loại được cho là đơn giản nhất để phân loại các dạng văn bản. Việc phân loại nội dung câu văn dựa vào các luật ngữ pháp tiếng việt. Các luật này có được là do nghiên cứu và đề xuất từ các chuyên gia. Đối với cách phân loại này, một loạt các biểu thức được tạo ra để so sánh với các nhãn từ đó đưa ra quyết định phân loại nội dung văn bản và nhãn của văn bản.

Tiếp cận dựa trên máy học là cách tiếp cận được sử dụng phổ biến rộng rãi để giải quyết bài toán phân loại nội dung văn bản. Cách tiếp cận này sẽ thay thế các kiến

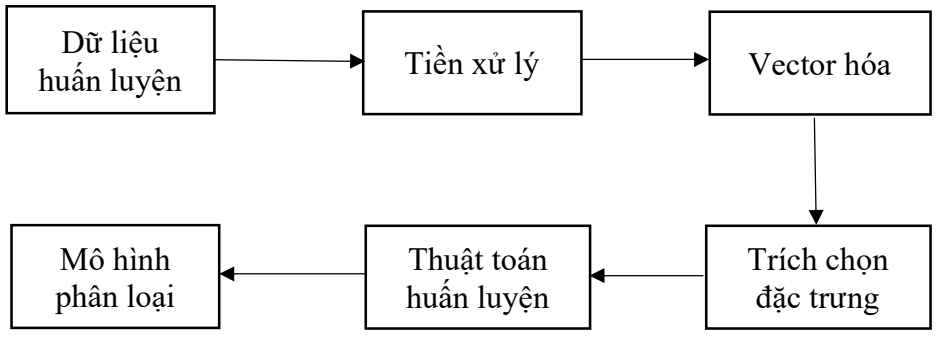
thức chuyên môn bằng một tập dữ liệu lớn các nội dung tiêu đề đã được gán nhãn (tập dữ liệu mẫu).

Cách tiếp cận dựa trên học máy được chia làm hai nhóm là nhóm các phương pháp học máy truyền thống và nhóm các phương pháp sử dụng mạng nơ-ron (Neural NetWork). Nhóm các phương pháp học máy truyền thống thường được sử dụng như là tính xác suất Naïve Bayes, Maximum Entropy, Máy Vector hỗ trợ (Support Vector machine - SVM),... Cách tiếp cận bằng học máy đã giải quyết được các hạn chế trong cách tiếp cận dựa trên luật. [12]

1.4.1 Phương pháp học máy truyền thống

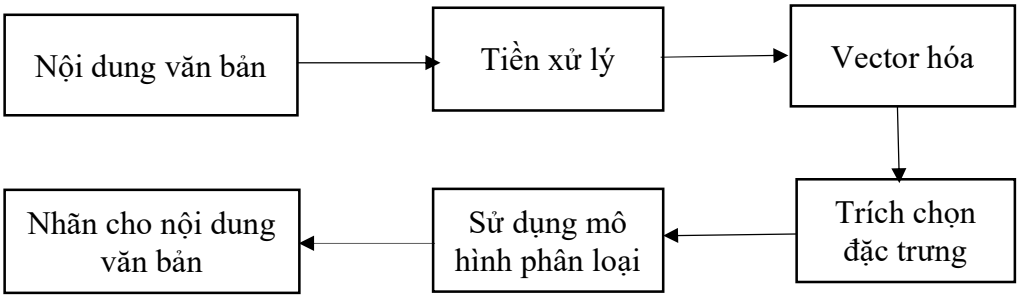
Với các phương pháp học máy truyền thống như SVM, cây quyết định,.. thì quá trình phân loại dữ liệu văn bản thường bao gồm hai giai đoạn sau:

- Giai đoạn huấn luyện: Là việc huấn luyện nhận đầu vào là tập các dữ liệu huấn luyện bao gồm các nội dung văn bản đã được gắn nhãn, sau khi xửa lý tập dữ liệu và áp dụng các thuật toán huấn luyện sẽ cho ra đầu ra là một mô hình phân loại.



Hình 1.4: Mô hình giai đoạn huấn luyện

- Giai đoạn phân lớp: là giai đoạn nhận đầu vào là nội dung tiêu đề của người dùng dưới dạng ngôn ngữ tự nhiên, sau quá trình tiền xử lý và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại tương ứng với từng nội dung của văn bản.



Hình 1.5: Mô hình giai đoạn phân lớp

❖ **Mô hình SVM**

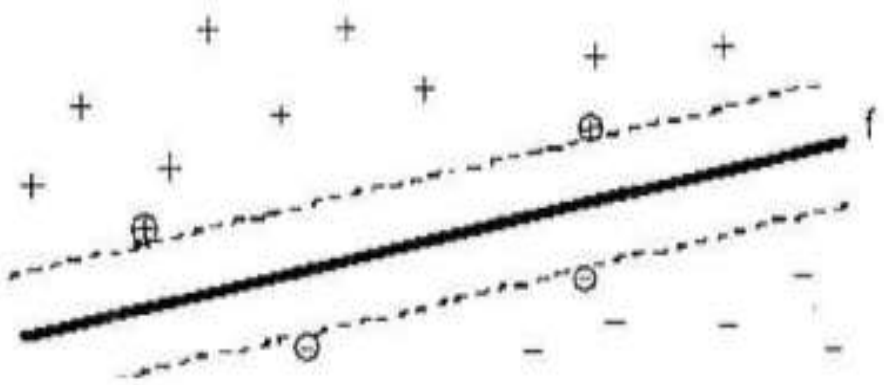
Giải thuật máy học vector hỗ trợ SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng năm 1995. Đây là một giải thuật phân lớp phổ biến, có hiệu quả cao và đã được áp dụng nhiều trong lĩnh vực khai phá dữ liệu và học máy.

Giải thuật SVM là giải thuật học máy có giám sát và được sử dụng trong các vấn đề phân lớp và hồi quy, chủ yếu là các bài toán phân lớp. SVM là một thuật toán phân loại nhị phân nhận dữ liệu đầu vào và phân loại chúng thành hai loại khác nhau. Với bộ dữ liệu huấn luyện thuộc hai loại cho trước, thuật toán huấn luyện SVM được xây dựng một mô hình SVM để phân loại các dữ liệu khác vào hai thể loại đó. [13]

Phương pháp này thực hiện phân lớp dựa trên các nguyên lý rủi ro thấp có cấu trúc SRM (Structural Risk Minimization), được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi. Các hàm công cụ SVM cho phép tạo không gian chuyển đổi để xây dựng các mặt phẳng phân lớp để tách các lớp ra thành các thành phần riêng biệt.

Giải thuật sẽ cho trước một tập dữ liệu huấn luyện bao gồm dữ liệu cùng với nhãn của chúng và được biểu diễn trong không gian vector, trong đó mỗi dữ liệu là một điểm, phương pháp này là tìm ra một mặt phẳng quyết định tốt nhất có thể và chia ra các điểm trong không gian thành hai lớp riêng biệt, tương ứng với lớp (+) và lớp (-). Chất lượng của mặt phẳng được quyết định bởi khoảng cách các điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng. Khi đó mặt phẳng biên càng lớn thì mặt phẳng quyết định càng tốt và việc phân loại càng chính xác.

Mục tiêu của phương pháp SVM là tìm ra được khoảng cách biên lớn nhất, điều này được minh họa như sau:



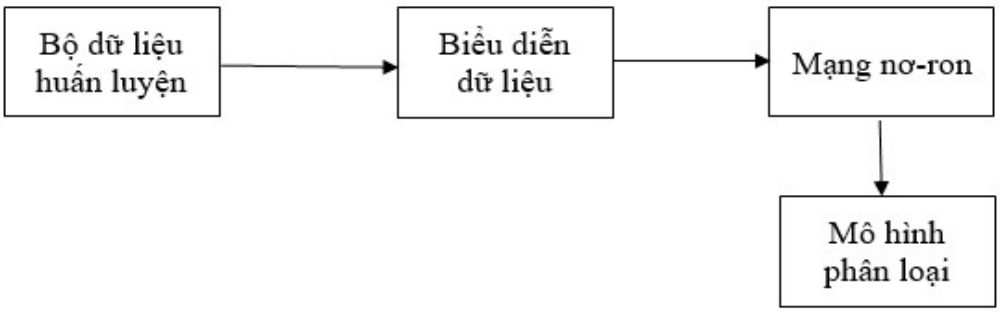
Hình 1.6: Mặt phẳng phân chia dữ liệu học thành 2 lớp (+) và lớp (-)

Đây là mô hình phổ biến và chính xác nhất trong một số các mô hình nổi tiếng về phân lớp dữ liệu.

1.4.2 Phương pháp sử dụng mạng nơ-ron

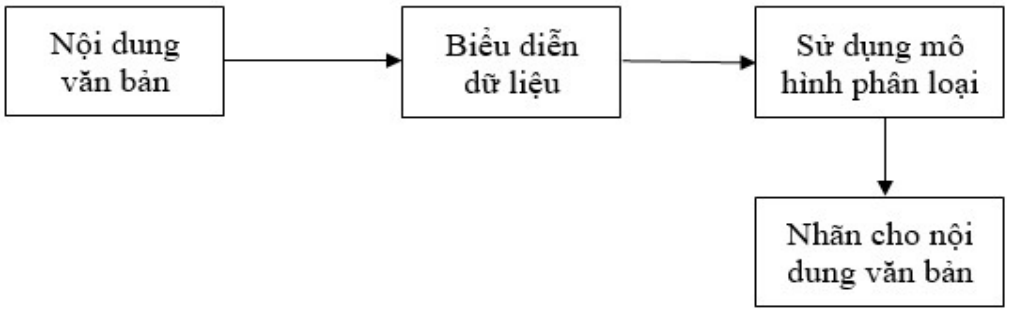
Với các phương pháp sử dụng mạng nơ-ron như Bert, RoBerta, PhoBert, v.v… thì quá trình phân loại dữ liệu văn bản cũng gồm có hai giai đoạn:

- Giai đoạn huấn luyện: Là giai đoạn đầu vào có tập dữ liệu huấn luyện bao gồm các nội dung tiêu đề đã được gắn nhãn, sau khi biểu diễn dữ liệu và đưa vào mạng nơ-ron sẽ cho ra đầu ra là một mô hình phân loại.



Hình 1.7: Mô hình giai đoạn huấn luyện sử dụng mạng nơ-ron

- Giai đoạn phân lớp: Là giai đoạn phân lớp đầu vào là nội dung tiêu đề của người dùng yêu cầu dưới dạng ngôn ngữ tự nhiên, sau quá trình biểu diễn dữ liệu và áp dụng mô hình phân loại sẽ cho ra nhãn phân loại của nội dung tiêu đề.



Hình 1.8: Mô hình giai đoạn phân lớp sử dụng mạng nơ-ron

❖ **Mô hình BERT**

BERT (Bidirectional Encoder Representations from Transformers) được hiểu là một mô hình được huấn luyện trước hay còn gọi là pre-train model, học các vector đại diện theo ngữ cảnh hai chiều của từ, được sử dụng để chuyển sang các bài toán khác trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Các kỹ thuật phổ biến như Word2vec, FastText hay Glove cũng tìm ra đại diện của từ thông qua ngữ cảnh chung của chúng. Tuy nhiên, những ngữ cảnh của các kỹ thuật này là đa dạng phong phú trong dữ liệu tự nhiên. Ví dụ các từ như "con chuột" có nghĩa khác nhau ở các ngữ cảnh khác nhau như "Con chuột máy tính này thật đẹp!" và "Con chuột này thật to". Trong khi các mô hình như Word2vec, fastText tìm ra một vector đại diện cho mỗi từ dựa trên một tập dữ liệu lớn nên không thể hiện được sự đa dạng của ngữ cảnh. Việc biểu diễn mỗi từ dựa vào các từ khác nhau trong câu thành một đại diện sẽ mang lại kết quả ý nghĩa rất nhiều.

Mô hình Bert đã tạo các biểu diễn theo ngữ cảnh dựa trên các từ trước và sau đó để dẫn đến một mô mình ngôn ngữ với ngữ nghĩa phong phú hơn. Điều này cho thấy mô hình Bert mở rộng khả năng của các phương pháp trước đây.

Các mô hình ngôn ngữ dựa trên LSTM (Long Short Term Memory) hai chiều đào tạo một mô hình ngôn ngữ tiêu chuẩn từ trái sang phải và cũng đào tạo một mô

hình ngôn ngữ từ phải sang trái (đảo ngược) dự đoán các từ trước, các từ tiếp theo. Sự khác biệt quan trọng là không LSTM nào đưa cả hai mã thông báo trước và sau vào cùng một lúc. [14]

CHƯƠNG 2: PHÂN TÍCH THIẾT KẾ ỨNG DỤNG

Chương 2 tập trung vào thiết kế các phương pháp phân loại văn bản theo dạng chủ đề, dùng mô hình phân tích để xử lý các chủ đề và đưa ra kết quả phân loại theo từng nhóm của chủ đề.

2.1 Sơ lược về phân loại nội dung tiêu đề trong mô hình OTT

Phân loại tên của chương trình phát sóng truyền hình có thể được quy đổi về bài toán lớn hơn là phân loại văn bản, phân loại câu văn hay từ vựng. Đây là các bài toán cơ bản về Xử lý Ngôn ngữ Tự nhiên (NLP Natural Language Processing). Bài toán phân loại tên chương trình được mô hình hóa qua mạng học sâu (deeplearning) với mô hình chuyển đổi giữa các câu văn (sequence-to-sequence Model). Dữ liệu đầu vào được gán nhãn và mô hình sẽ học từ dữ liệu được gán nhãn cho trước, sau đó sẽ được dùng để dự đoán các nhãn tương ứng cho các dữ liệu mới trong mô hình.

Phân loại tên chương trình truyền hình có thể được định nghĩa như sau. Từ một tập các văn bản $D = \{d_1, d_2, \ldots, d_n\}$, được gọi là tập huấn luyện, trong đó các tên chương trình truyền hình được gán nhãn chủ đề c_i với c_i thuộc tập các tiêu đề $C = \{c_1, c_2, \ldots, c_n\}$ để xây dựng bộ phân loại. Nhiệm vụ của bộ phân loại là gán đúng nhãn tiêu đề c_k cho một tên chương trình mới thuộc d_k bất kỳ, trong đó c_k thuộc vào tập tiêu đề C. [15]

Phân loại tên chương trình đã thu hút rất nhiều các nhà nghiên cứu và đạt được nhiều thành công đặc biệt là đối với ngôn ngữ tiếng Anh. Tên chương trình có thể được phân loại dựa trên nhiều hướng tiếp cận khác nhau như kỹ thuật máy học, phân cụm hoặc luật kết hợp. Trong số các hướng tiếp cận trên thì hướng tiếp cận sử dụng máy học như là bộ phân loại thu hút được nhiều nhà nghiên cứu nhất và cho kết quả khả quan. Một số kỹ thuật thường được sử dụng là: SVM, Bert, PhoBert, v.v….

Phân loại tên chương trình có thể dựa trên mô hình BERT (Biderectional Encoder Representations from Transformers), là một dạng mô hình mới của Google AI cho NLP. BERT dùng thông tin về ngôn ngữ được học trước để xử lý các bài toán như thiết bị tự động trả lời, phân tích cảm xúc câu trả lời, tìm ý chính của đoạn văn.

Hai phiên bản PhoBERT là “base” và “large” là mô hình ngôn ngữ quy mô lớn đầu tiên được đào tạo trước cho tiếng việt. Phương pháp tiếp cận đào tạo trước của PhoBERT dựa trên RoBERTa, tối ưu hóa quy trình đào tạo của BERT để có hiệu suất chính xác hơn.

PhoBERT vượt trội hơn so với các phương pháp tiếp cận đơn ngữ và đa ngôn ngữ trước đây, mô hình đã đạt được những kết quả tốt nhất về các nhiệm vụ xử lý ngôn ngữ của Việt Nam.

2.2 Quy trình phân loại nội dung tiêu đề trong mô hình OTT

Bộ dữ liệu được lấy dữ liệu từ trang web lịch phát sóng VTV, loại bỏ các thẻ HTML, JavaScript, … để có bộ dữ liệu tốt và cho kết quả xử lý dữ liệu chính xác.

Thực hiện tách từ là một công đoạn quan trọng nhất trong xử lý ngôn ngữ tự nhiên, do Tiếng Việt có độ phức tạp cao hơn ngôn ngữ khác (bởi có các từ ghép). Việc tách từ theo nhiều cách khác nhau có thể gây ra sự hiểu nhầm về mặt ngữ nghĩa. Tuy nhiên, có một số công cụ hỗ trợ thực hiện việc này, phổ biến nhất là VnTokenizer. Chuẩn hóa từ để đưa tiêu đề từ các dạng không đồng nhất về cùng một dạng (Ví dụ như tất cả quy định về chữ thường).

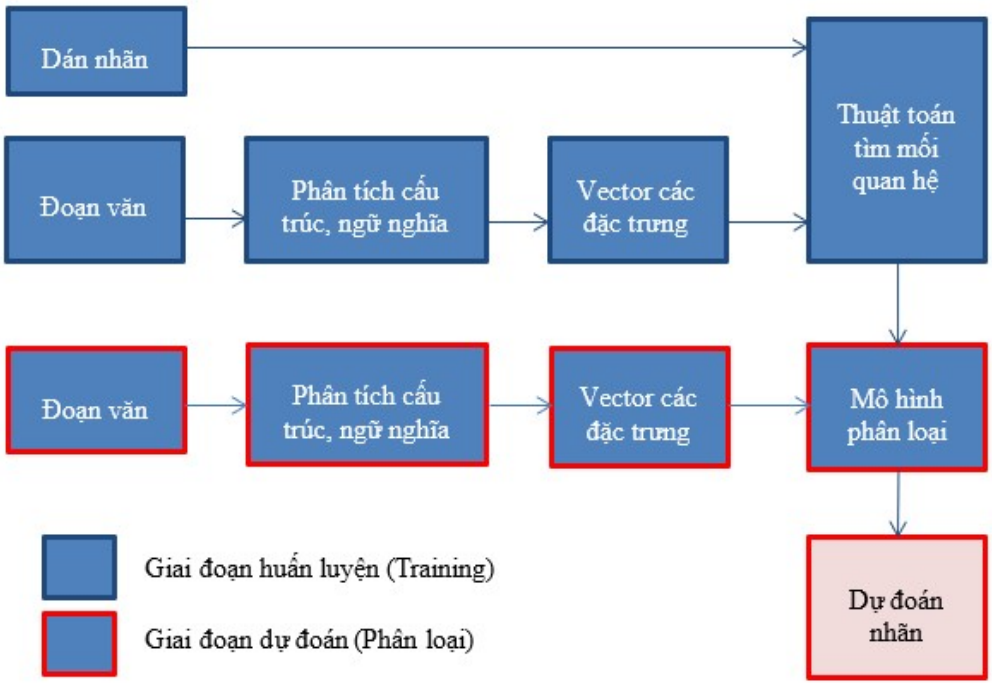
Việc tối ưu hóa bộ nhớ lưu trữ và tính chính xác rất quan trọng. Có nhiều cách viết, mỗi cách viết khi lưu trữ sẽ tốn dung lượng bộ nhớ khác nhau. Do đó, tùy theo nhu cầu, tình hình thực tế để đưa ra tiêu đề về một dạng đồng nhất.

Bước trích xuất đặc trưng gồm 2 bước là xây dựng bộ từ điển và tạo vector số cho các nội dung tiêu đề theo phương pháp túi đựng từ (Bag of word - BoW). Tất cả các từ trong nội dung tiêu đề cần được chuyển thành dạng biểu diễn số. Sau đó sẽ thay thế từ đó bằng thứ tự xuất hiện trong bộ từ điển và tiến hành xây dựng từ điển chứa tất cả các từ trong tập dữ liệu sau khi đã tiến hành tách từ và loại bỏ stop words. Cuối cùng sẽ thu được vector thuộc tính cho từng tập tin trong tập dữ liệu. Mỗi vector sẽ có độ dài bằng số từ trong từ điển.

Bước xây dựng mô hình các thuật toán học máy sẽ huấn luyện một bộ phân loại sử dụng các vector thuộc tính của dữ liệu ở trên. Có nhiều mô hình học máy có thể được sử dụng để huấn luyện tạo ra mô hình cuối cùng. Trong nghiên cứu này đã

sử dụng mô hình PhoBert để huấn luyện bao gồm lớp đầu vào, các lớp ẩn và lớp đầu ra.

Mô hình phân loại dữ liệu gồm hai giai đoạn:



Hình 2.1: Mô hình phân loại văn bản

- Giai đoạn huấn luyện (training) là giai đoạn học tập của mô hình phân loại văn bản. Trong bước này, mô hình sẽ học từ dữ liệu có nhãn. Dữ liệu văn bản sẽ được số hóa thông qua bộ trích xuất đặc trưng (feature extractor) để mỗi mẫu dữ liệu trong tập huấn luyện trở thành một vector nhiều chiều (đặc trưng). Thuật toán máy học sẽ học và tối ưu các tham số để đạt được kết quả tốt trên tập dữ liệu. Nhãn của dữ liệu được dùng để đánh giá việc mô hình học tốt không và dựa vào đó để tối ưu.
- Giai dự đoán (prediction), là giai đoạn sử dụng mô hình học máy sau khi nó đã học xong. Ở giai đoạn này, dữ liệu cần dự đoán cũng vẫn thực hiện các bước trích xuất đặc trưng. Mô hình đã học sau đó nhận đầu vào là đặc trưng đó và đưa ra kết quả dự đoán. [16]

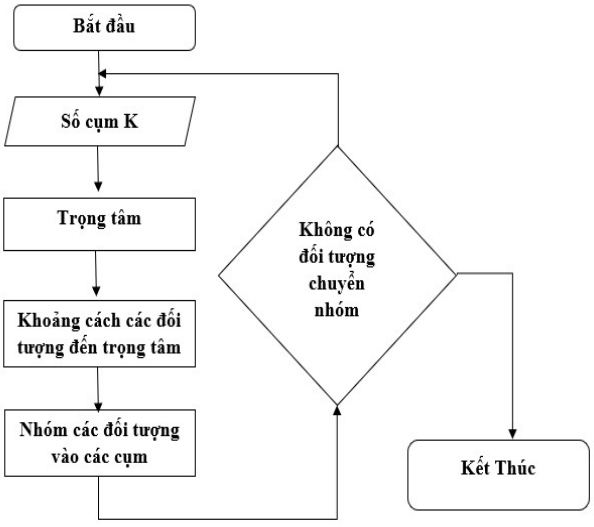
2.3 Thuật toán K-Means

2.3.1 Giới thiệu về K-Means

K-means là thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm K-Means là phân chia một bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý khác nhau. Các điểm dữ liệu trong cùng một cụm thì phải có cùng một số tính chất nhất định. Tức là giữa các điểm trong cùng một cụm phải có sự liên kết lẫn nhau. Đối với máy tính thì các điểm trong một cụm đó sẽ là các điểm dữ liệu gần nhau.

Thuật toán K-Means là một trong những phương pháp sử dụng trong phân tích tính chất phân cụm của dữ liệu. Thuật toán K-Means đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành nhiều nhóm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta biết nó thuộc về nhóm nào. [17]

2.3.2 Các bước của thuật toán K-Means



Hình 2.2: Sơ đồ thuật toán K-Means

Xây dựng bộ dữ liệu được thực hiện qua các giai đoạn theo sơ đồ thuật toán K-Means ở Hình 2.2.

Đầu vào: Dữ liệu X và số lượng cụm cần tìm K.

Đầu ra: Các điểm trọng tâm M và nhãn vector cho từng điểm dữ liệu Y.

- Chọn K điểm bất kỳ làm các điểm trọng tâm ban đầu.
- Phân mỗi điểm dữ liệu vào các cụm có điểm trọng tâm gần nó nhất.
- Nếu việc gán dữ liệu vào từng cụm ở bước 2 không thay đổi so với vòng lặp trước đó thì ta dừng thuật toán.
- Cập nhật lại điểm trọng tâm cho từng cụm bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cụm đó sau bước 2.
- Quay lại bước 2. [18]

2.3.3 Ưu và nhược điểm của thuật toán K-Means

❖ **Ưu điểm:**

Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.

Luôn có số phần tử K cụm dữ liệu và luôn có ít nhất một điểm dữ liệu trong một cụm dữ liệu.

Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau

Mỗi thành phần của một cụm là gần với chính cụm đó hơn là bất cứ một cụm nào khác.

❖ **Nhược điểm:**

Không có khả năng tìm ra các cụm không lồi hoặc các cụm có hình dạng phức tạp.

Khó khăn trong việc xác định trọng tâm của các cụm ban đầu vì vậy chọn ngẫu nhiên các trọng tâm cụm lúc khởi tạo, độ hội tụ của thuật toán phụ thuộc vào việc khởi tạo các vector trung tâm cụm.

Khó để chọn ra được số lượng cụm tối ưu ngay từ đầu, mà phải qua nhiều lần thử để tìm ra được số lượng cụm tối ưu.

Rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu

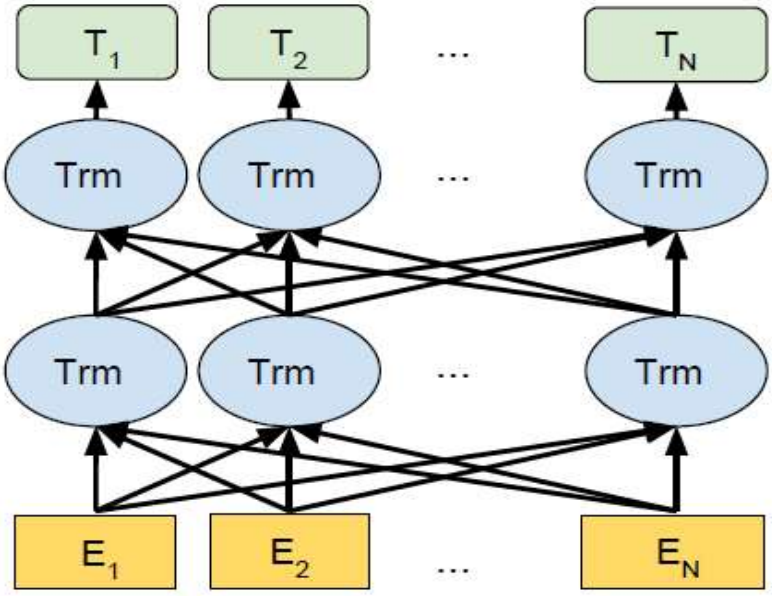
Không phải lúc nào mỗi đối tượng cũng chỉ thuộc về một cụm, chỉ phù hợp với đường biên giữa các cụm rõ. [19]

2.4 Giới thiệu mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) (tạm dịch: Mô hình mã hóa hai chiều dữ liệu từ các khối Transformer), là một phương pháp kỹ thuật được xây dựng dựa trên mô hình mạng mô phỏng theo hệ thống nơ-ron thần kinh của con người (neural network) dùng để đào tạo trước (pre-train) quá trình xử lý ngôn ngữ tự nhiên. Nói một cách đơn giản, thì nó có thể được sử dụng để giúp Google phân biệt rõ hơn ngữ cảnh của các từ xuất hiện trong truy vấn tìm kiếm.

Điểm đột phá của BERT nằm ở khả năng huấn luyện các mô hình ngôn ngữ dựa trên toàn bộ tổ hợp các từ trong một câu hoặc truy vấn (huấn luyện hai chiều), thay vì cách thức huấn luyện truyền thống dựa trên thứ tự xuất hiện của các từ (từ trái qua phải hoặc kết hợp giữa trái qua phải và phải qua trái). BERT cho phép mô hình ngôn ngữ học về ngữ cảnh của từ vựng dựa trên các từ xung quanh nó, thay vì chỉ dựa vào từ ngữ đứng trước hoặc ngay sau nó.

Kiến trúc mô hình BERT là một bộ mã hóa Transformer hai chiều (bidirectional Transformer encoder). Việc sử dụng Transformer không có gì đáng ngạc nhiên vì đây là một xu hướng gần đây do tính hiệu quả và hiệu suất vượt trội của huấn luyện Transformers trong việc phát hiện các phụ thuộc với khoảng cách xa (long-distance dependencies) so với kiến trúc Recurrent neural network. Trong khi đó, bộ mã hóa hai chiều (bidirectional encoder) là một tính năng nổi bật giúp phân biệt BERT với OpenAI GPT (sử dụng từ trái sang phải Transformer) và kết hợp giữa huấn luyện từ trái sang phải và một mạng riêng rẽ từ phải sang trái LSTM. [20]. Kiến trúc của mô hình Bert được mô tả theo sơ đồ Hình 2.3.



Hình 2.3: Kiến trúc của mô hình BERT

Sử dụng bộ mã hóa Transformer đã được huấn luyện, BERT có thể biểu diễn bất kỳ token nào dựa trên ngữ cảnh hai chiều của nó. Trong quá trình học có giám sát trên các tác vụ, BERT tương tự như GPT ở hai khía cạnh.

- Đầu tiên, các biểu diễn BERT sẽ được truyền vào một tầng đầu ra được bổ sung, với những thay đổi tối thiểu tới kiến trúc mô hình tùy thuộc vào bản chất của tác vụ, chẳng hạn như dự đoán cho mỗi token hay dự đoán cho toàn bộ chuỗi.
- Thứ hai, tất cả các tham số của bộ mã hóa Transformer đã được huấn luyện đều được tinh chỉnh, trong khi tầng đầu ra bổ sung sẽ được huấn luyện từ đầu.

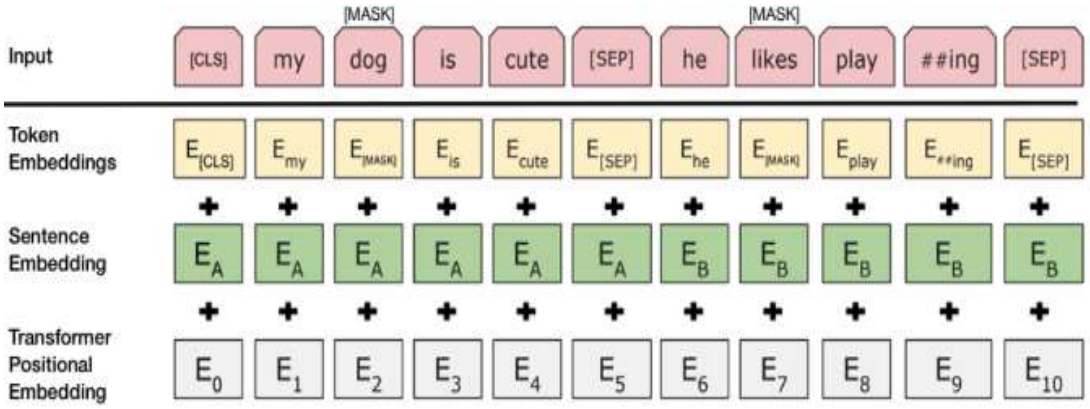
2.4.1 Biểu diễn đầu vào của Bert

Đầu vào có thể biểu diễn của một câu văn bản đơn hoặc một cặp câu văn bản. (Ví dụ: [Câu hỏi, câu trả lời]) được đặt thành một chuỗi tạo bởi các từ.

Chuỗi đầu vào BERT biểu diễn một cách tường minh cả văn bản đơn và cặp văn bản. Với văn bản đơn, chuỗi đầu vào BERT là sự ghép nối của token phân loại đặc biệt “<cls>”, token của chuỗi văn bản, và token phân tách đặc biệt “<sep>”. Với cặp văn bản, chuỗi đầu vào BERT là sự ghép nối của “<cls>”, token của chuỗi văn

bản đầu, “<sep>”, token của chuỗi văn bản thứ hai, và “<sep>”. Ta sẽ phân biệt nhất quán thuật ngữ “chuỗi đầu vào BERT” với các kiểu “chuỗi” khác. Chẳng hạn, một chuỗi đầu vào BERT có thể bao gồm cả một chuỗi văn bản hoặc hai chuỗi văn bản.

Khi có một chuỗi đầu vào cụ thể được xây dựng bằng cách tính tổng các token đó với vector phân đoạn và vị trí tương ứng của các từ trong chuỗi. Cho dễ hình dung, biểu diễn đầu vào được trực quan hóa trong hình dưới đây:



Hình 2.4: Mô hình đại diện đầu vào của BERT

Token đầu tiên cho mỗi chuỗi được mặc định là một token đặt biệt có giá trị là [CLS]. Đầu ra của Transformer tương ứng với token này sẽ là được sử dụng để đại diện cho cả câu trong các nhiệm vụ phân loại nhãn chương trình. Nếu không trong các nhiệm vụ phân loại vector này được bỏ qua.

Trong trường hợp các cặp câu được gộp lại với nhau thành một chuỗi duy nhất chúng ta phân biệt các câu theo hai cách. Đầu tiên, chúng ta tách chúng bởi một token đặt biệt [SEP]. Thứ hai, chúng ta thêm một segment embedding cho câu A và một segment embedding khác cho câu B.

Khi chỉ có một câu đơn duy nhất segement embedding chỉ có cho câu A.

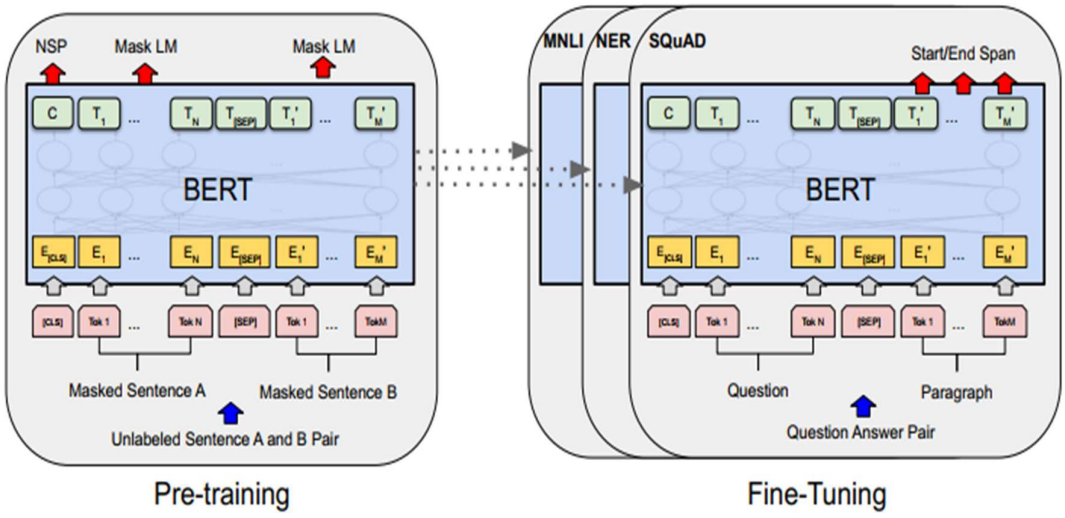
Kiến trúc hai chiều của BERT là bộ mã hóa Transformer. Thông thường trong bộ mã hóa Transformer, các embedding vị trí được cộng vào mỗi vị trí của chuỗi đầu vào BERT. Tuy nhiên, khác với bộ mã hóa Transformer nguyên bản, BERT sử dụng các embedding vị trí có thể học được cho thấy các embedding của chuỗi đầu vào BERT là tổng các embedding của token, embedding đoạn và embedding vị trí. [21]

2.4.2 Cải thiện BERT

BERT được training bằng cách sử dụng "Mask Language Model". Ý tưởng của phương pháp là khi ta đưa một nội dung vào mô hình thì 15% token của một nội dung sẽ được thay thế bằng token <mask> việc của mô hình là sẽ dự đoán từ được <mask>.

VD: Sau đó đám cháy đã lan nhanh sang các khu vực khác.

Sau đó <mask> cháy đã lan nhanh sang các khu vực khác.



Hình 2.5: Quy trình tổng thể pre-training và fine-tuning của BERT

Một kiến trúc được sử dụng cho cả pretrain-model và fine-tuning model. Chúng ta sử dụng cùng một tham số huấn luyện trước để khởi tạo mô hình cho các tác vụ downstream khác nhau. Trong suốt quá trình điều chỉnh thì toàn bộ các tham số của các lớp học chuyển giao sẽ được tinh chỉnh. [CLS] là một ký tự (token) đặc biệt được thêm vào trước mọi ví dụ đầu vào và [SEP] là một ký tự phân tách đặc biệt (ví dụ: phân tách các câu hỏi / câu trả lời). [22]

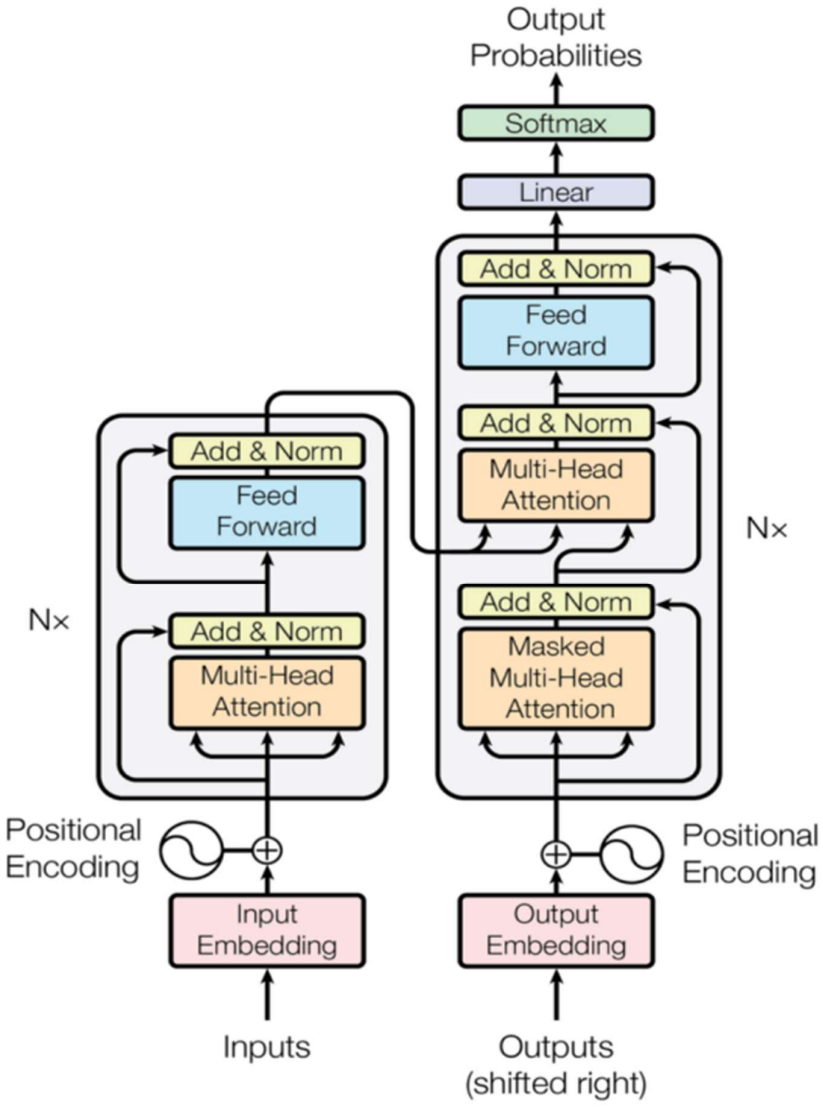
2.4.3 Pre-training BERT

Các mô hình ngôn ngữ truyền thống từ trái sang phải hoặc từ phải sang trái không được sử dụng để Pre-training BERT mà sử dụng hai nhiệm vụ không được giám sát (unsupervised tasks).

Nhiệm vụ số 1: Masked Language Modeling (MLM) là một tác vụ cho phép tinh chỉnh lại các biểu diễn từ trên các bộ dữ liệu không được giám sát bất kỳ. MLM có thể được áp dụng cho những ngôn ngữ khác nhau để tạo ra biểu diễn nhúng cho chúng.

Nhiệm vụ số 2: Next Sentence Prediction (NSP) là một bài toán phân loại học có giám sát với 2 nhãn. Đầu vào của mô hình là một cặp câu sao cho 50% câu thứ 2 được lựa chọn là câu tiếp theo của câu thứ nhất và 50% được lựa chọn một cách ngẫu nhiên từ bộ văn bản mà không có mối liên hệ gì với câu thứ nhất. Nhãn của mô hình sẽ tương ứng với IsNext khi cặp câu là liên tiếp hoặc NotNext nếu cặp câu không liên tiếp.

Cũng tương tự như mô hình Question and Answering, chúng ta cần đánh dấu các vị trí đầu câu thứ nhất bằng ký tự [CLS] và vị trí cuối các câu bằng ký tự [SEP]. Các ký tự này có tác dụng nhận biết các vị trí bắt đầu và kết thúc của từng câu thứ nhất và thứ hai. [23]

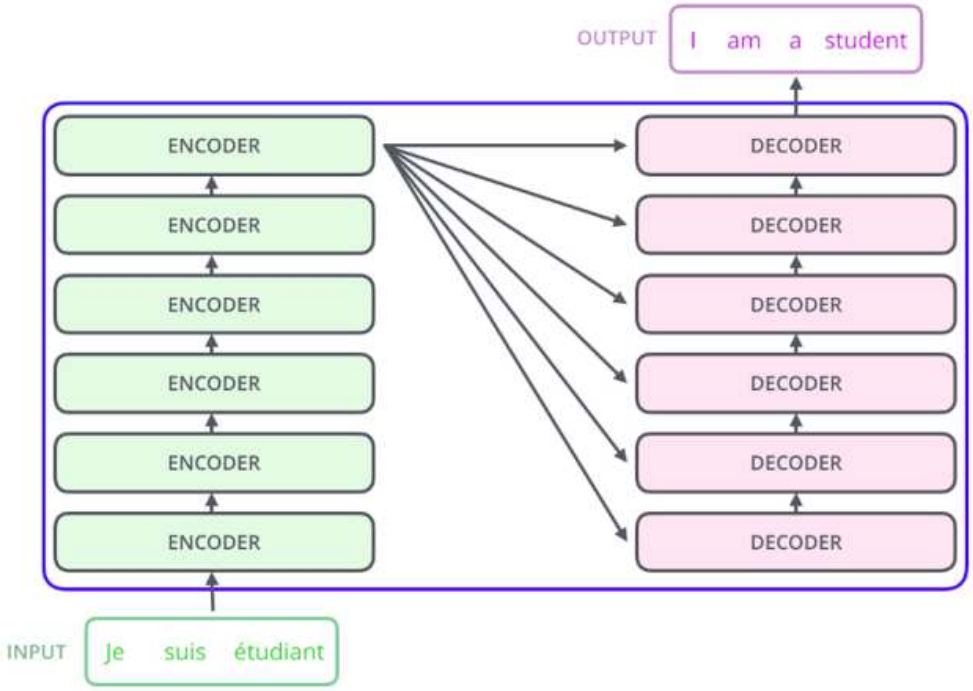


Hình 2.6: Sơ đồ kiến trúc mô hình BERT cho tác vụ NSP

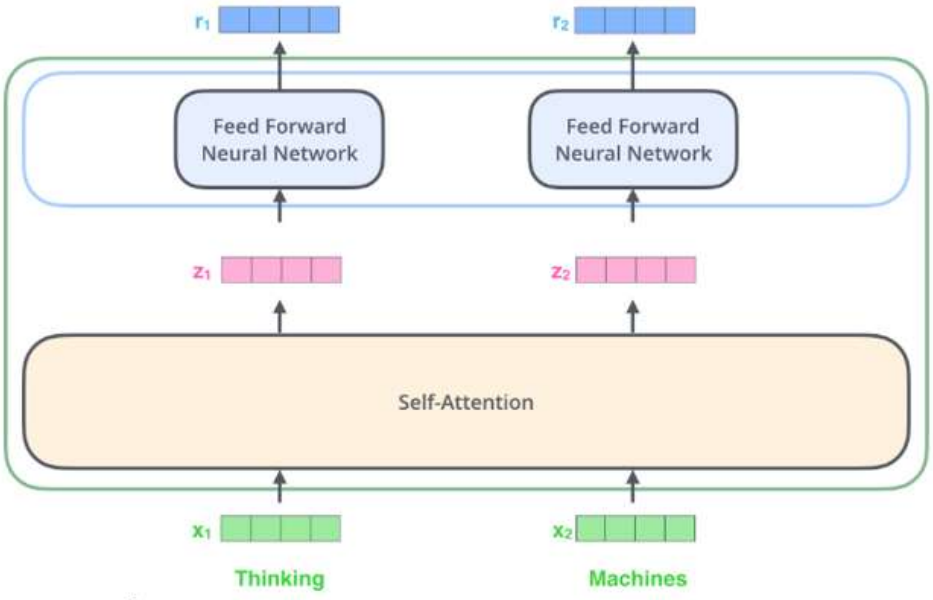
2.4.4 Kiến trúc của BERT

Transformers là mô hình sử dụng phương thức attention. Sự ra đời của transformer đã giải quyết được vấn đề của các mô hình như LSTM, BiLSTM là mất thông tin, thời gian training dữ liệu lâu. Phương thức attention hiểu đơn giản các từ quan trọng trong một câu sẽ có ảnh hưởng lớn hơn đến ý nghĩa của câu.

Transformers gồm 2 phần encoder và decoder: bao gồm nhiều block xếp chồng lên nhau. Với BERT chỉ sử dụng Encoder.

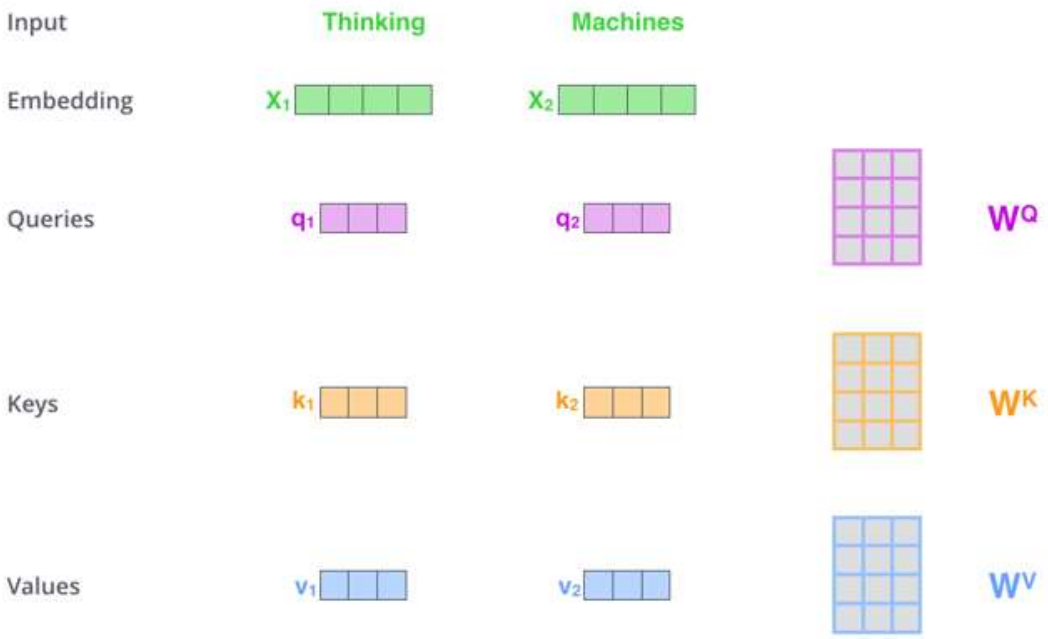


Hình 2.7: Kiến trúc transformer



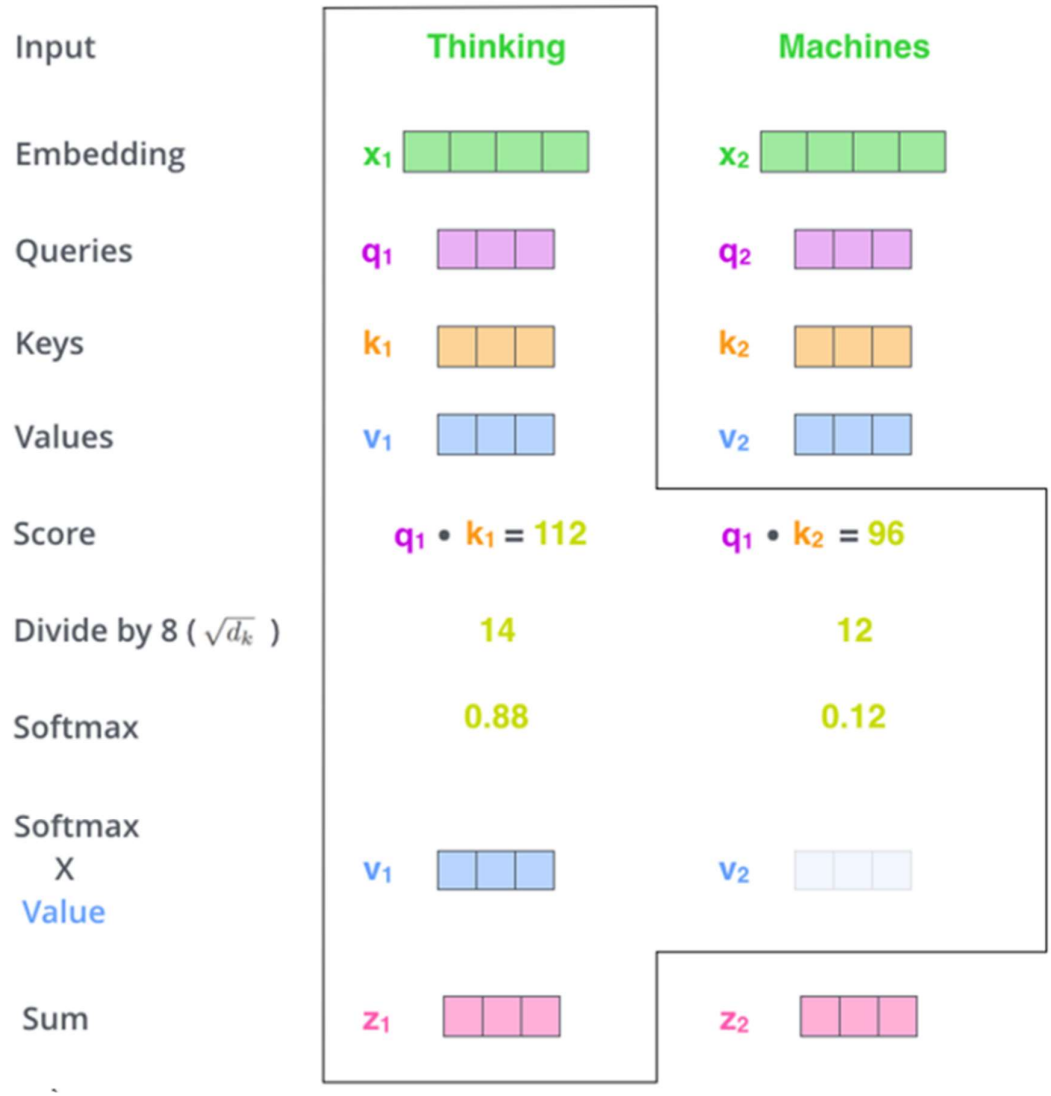
Hình 2.8: Kiến trúc của một block transformer

Các từ khi đưa vào sẽ được embedding về dạng vector (ở ví dụ trên ta có vector có kích thước là 1x64) được thu gọn lại để dễ biểu diễn. Tiếp theo sẽ đến bước tính toán.



Hình 2.9: Mô hình kiến trúc Self-Attention

Ta có 3 ma trận tham số Wq, Wk, Wv đại diện cho Querry, Key, Value sẽ được khởi tạo có kích thước số hàng bằng với chiều của vector, số cột tùy thuộc vào cách ta chọn. Ta lần lượt lấy các vector của từ nhân với các ma trận này ta sẽ thu được các vector tương ứng như trên Hình 2.9 (Tích ma trận)



Hình 2.10: Mô hình tính một vector Attention

Tiếp đó ta sẽ lần lượt lấy vector q nhân tích vô hướng với các vector k trong câu ta sẽ thu được trọng số. Các trọng số được chia cho căn bậc 2 chiều của vector (ở đây vector có chiều dài là 64) sau đó đưa qua hàm softmax. Làm lần lượt tương tự với các từ còn lại trong câu ta sẽ thu được các vector z còn lại. Đó là kiến trúc của 1 block encoder transformer.

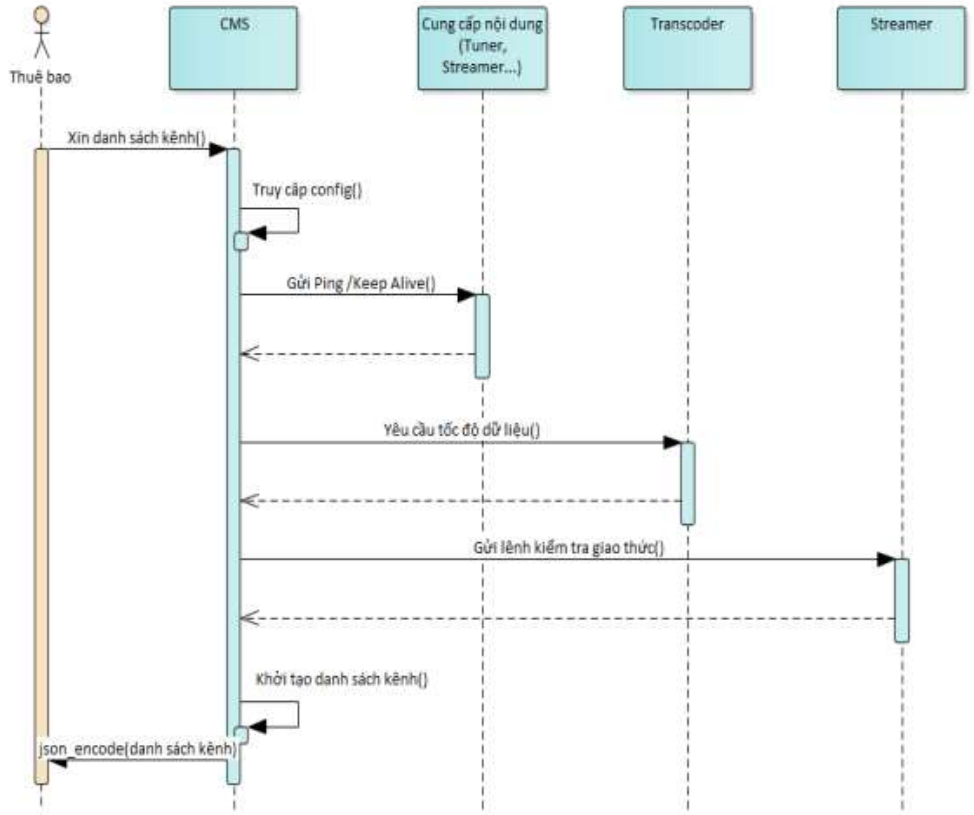
Như vậy việc xếp chồng nhiều block thì output của 1 block sẽ là input của block khác. (Vector z sẽ có cùng kích thước với vector đầu vào). Mô hình BERT sẽ có 12 block xếp chồng lên nhau, hay còn gọi là 12 lớp. Với BERT large ta sẽ có 24 block. [24]

CHƯƠNG 3: TRIỂN KHAI ỨNG DỤNG

Trong chương này trình bày các phương án cài đặt phương pháp tự động phân loại và bổ sung theo từng chủ đề của chương trình Truyền hình đã được xây dựng ở Chương 2.

3.1 Sơ đồ chức năng hiển thị danh sách kênh

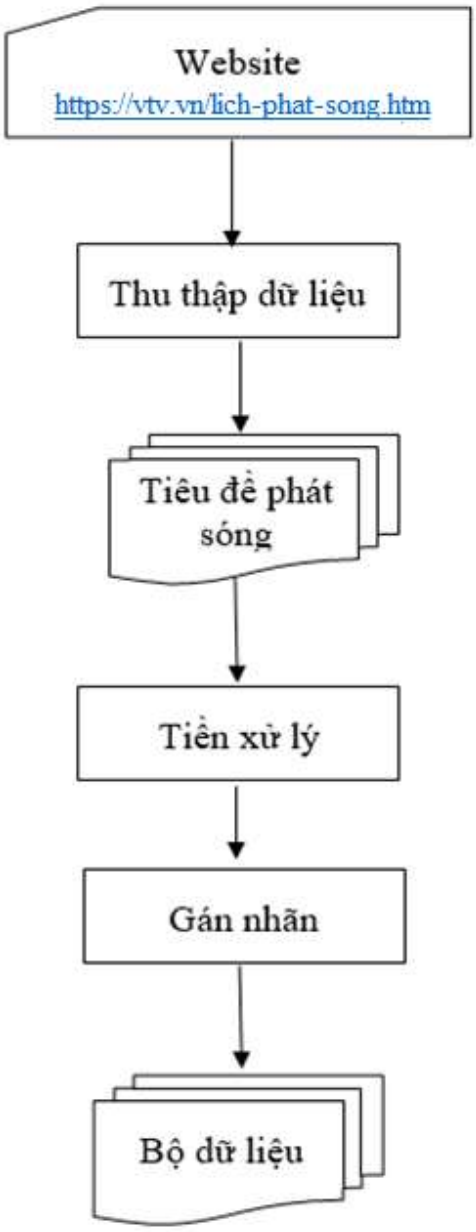
Với mô hình OTT thì khi người dùng muốn xem danh sách các chương trình có thể xem vào các thời điểm nào đó. Người dùng có thể nhập tên một thể loại bất kỳ thì đầu ra của danh sách sẽ khởi tạo một danh sách chương trình phát sóng theo thể loại của người dùng yêu cầu, không lệ thuộc vào khái niệm "Kênh" cổ điển, chỉ tập trung vào nội dung / chủ đề của nội dung mà người thuê bao quan tâm. Hình 3.1 thể hiện sơ đồ chức năng hiển thị danh sách các kênh cho người dùng theo từng chủ đề.



Hình 3.1: Sơ đồ chức năng cập nhật danh sách kênh cho người dùng

3.2 Xây dựng bộ dữ liệu

Việc thực hiện xây dựng bộ dữ liệu luận văn đã thực hiện theo các giai đoạn trong mô hình dưới đây:



Hình 3.2: **Mô hình xây dựng bộ dữ liệu**

3.2.1 Thu thập dữ liệu

Luận văn được lấy dữ liệu phát sóng từ trang web VTV: https://vtv.vn/lich-phat-song.htm

Dữ liệu khoảng 1000 tiêu đề phát sóng của dịch vụ truyền hình VTV mỗi ngày. Nội dung bao gồm các chủ đề như: phim truyện, thời sự, thể thao, giải trí, ca nhạc, kỹ năng sống, trẻ em, du lịch. [25]

Dữ liệu được lấy từ trang web theo định dạng như Hình 3.3.

Hình 3.3: Dữ liệu từ trang web VTV

3.2.2 Tiền xử lý

Bộ dữ liệu sau khi thu thập được từ trang web VTV sẽ được tiến hành xử lý. Luận văn thực hiện tiền xử lý dữ liệu bằng cách loại bỏ các thẻ HTML, JavaScript….

Cài đặt xử lý dữ liệu trên ngôn ngữ Python như sau:

```
from bs4 import BeautifulSoup
import requests
import pandas as pd
from datetime import date
import os
VTV_link = "https://vtv.vn/lich-phat-song.htm"
def crawler():
    today = date.today()
    file_name = 'data_' + today.strftime("%d_%m_%Y") + '.xlsx'
    response = requests.get(VTV_link)
    soup = BeautifulSoup(response.text, "html.parser")
    title = soup.findAll("span", class_='title')
```

```
    ls_title = [i.text for i in title]
    data_frame = {"title":ls_title}
    data_frame = pd.DataFrame(data=data_frame)
    data_frame.to_excel(os.path.join('data', file_name))
if __name__=='__main__':
    crawler()
```

3.2.3 Gán nhãn

Bộ dữ liệu thuộc về lĩnh vực dịch vụ phát sóng truyền hình nên luận văn được tiến hành gán nhãn cho bộ dữ liệu theo các chủ đề như: Giải trí, Du lịch, Phim truyện, Ca nhạc, Thể thao, Thời sự, Kỹ năng sống, Trẻ em.

- **Sử dụng thuật toán K-Means để gán nhãn cho bộ dữ liệu:**

Cài đặt thuật toán K-Means trên ngôn ngữ Python:

```
import pandas as pd
from collections import Counter
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import argparse
from tqdm import tqdm
map_dict_label = {
    2: 'phim truyện',
    3: 'ca nhạc',
    0: 'giải trí',
    5: 'thời sự',
    6: 'kĩ năng sống',
    4: 'thể thao',
    1: 'du lịch',
    7: 'trẻ em'
}
if __name__=='__main__':
```

```
arg_parser = argparse.ArgumentParser()
arg_parser.add_argument('--data_dir')
args = arg_parser.parse_args()
data = pd.read_excel(args.data_dir)
all_text, all_label_text = [], []
for i in tqdm(range(data.shape[0])):
    title = data['title'][i].lower()
    all_text.append(title)
vectorizer=TfidfVectorizer()
X = vectorizer.fit_transform(all_text)
print(X.shape)
cluster = KMeans(n_clusters=8)
labels = cluster.fit_predict(X)
text_label = [map_dict_label[i] for i in labels]
data_frame = {
    'title' : all_text,
    'label' : text_label
}
df = pd.DataFrame(data_frame, columns=['title','label'])
df.to_excel('data/labed_data.xlsx')
```

- **Tập nhãn luận văn xây dựng bao gồm 8 nhãn:**

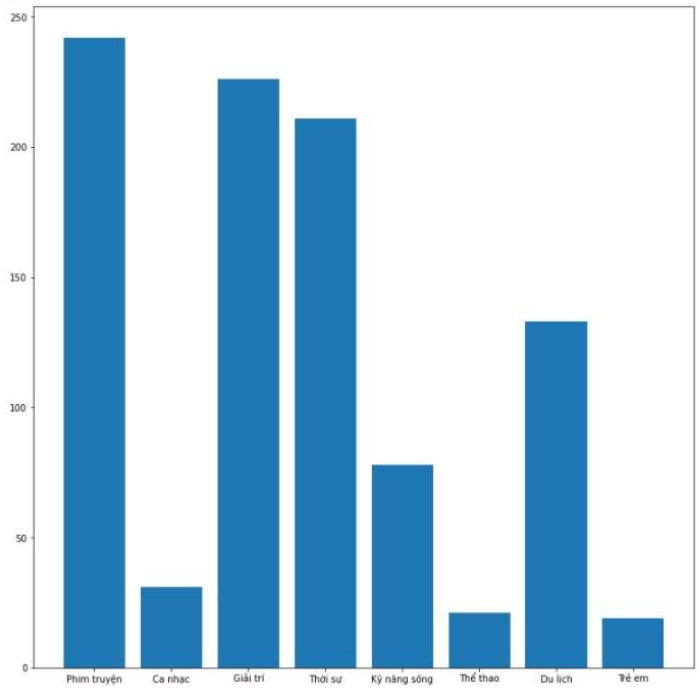
Bảng 3.1: Bảng nhãn và ví dụ

STT	Tiêu đề phát sóng truyền hình	Nhãn
1	Chuyến đi màu xanh	Phim truyện
2	Hành trang cuộc sống	Ca nhạc
3	Hành trình yêu thương	Kỹ năng sống
4	Không gian xanh - nhà tknt thanh thảo - nhà có thú cưng	Giải trí
5	Ký ức sài gòn - thành phố hồ chí minh	Thời sự
6	Sự kiện thể thao	Thể thao
7	Thế giới dưới nước	Trẻ em
8	Du lịch kỳ thú: gõ cửa thăm nhà	Du lịch

3.2.4 Thống kê bộ dữ liệu

Bảng 3.2: Thống kê tần suất các nhãn trong bộ dữ liệu

STT	Nhãn	Số lượng Tiêu đề	Tỉ lệ trong bộ dữ liệu (%)
1	Phim truyện	242	25,18
2	Ca nhạc	31	3,22
3	Kỹ năng sống	78	8,11
4	Giải trí	226	23,51
5	Thời sự	211	21,95
6	Thể thao	21	2,18
7	Trẻ em	19	1,97
8	Du lịch	133	13,83



Hình 3.4: Biểu đồ số lượng các nhãn của chương trình dùng để training

3.3 Thiết lập thực nghiệm

Với bộ dữ liệu chuẩn bị cho thiết lập thực nghiệm, luân văn lấy được 1000 tiêu đề của lịch phát sóng truyền hình theo pháp quy tiếng việt. Luận văn được chia thành 8 nhãn.

Để đánh giá kết quả của việc xác định thực thể và thuộc tính ta đánh giá thông qua độ chính xác (precision), độ bao phủ (recall), tính cân bằng của độ chính xác và độ bao phủ (F1) được xác định như sau: [26]

$$precision = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ được\ gán}$$

$$recall = \frac{số\ nhãn\ gán\ đúng}{tổng\ số\ nhãn\ thực\ tế}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

3.4 Công cụ thực nghiệm

Luận văn sử dụng hai công cụ thực nghiệm là sklearn svm Linear SVC sử dụng cho mô hình SVM và simpletransformers sử dụng cho hai mô hình còn lại là BERT multilingual và PHOBERT.

Sklearn svm Linear SVC

Sklearn svm Linear SVC tương tự như SVC với tham số kernel = "linear", nhưng được triển khai dưới dạng liblinear chứ không phải libsvm, nó linh hoạt hơn trong việc lựa chọn các hàm penalties và hàm loss và nên mở rộng quy mô tốt hơn đến số lượng lớn dữ liệu.

Để cài đặt công cụ ta dùng lệnh:

Pip install sklearn

Simpletransformer model được xây dựng với một nhiệm vụ xử lý ngôn ngữ tự nhiên cụ thể. Mỗi mô hình như vậy được trang bị các tính năng và chức năng được thiết kế để phù hợp nhất với nhiệm vụ mà chúng dự định thực hiện.

Để cài đặt sử dụng ta dùng lệnh:

Pip install simpletransformers

Pip install transformer

Pip install underthesea

Pip install torch

Pip install scikit-learn

Cả 3 mô hình đều sử dụng công cụ ngôn ngữ Python.

- **Cài đặt mô hình SVM trên ngôn ngữ Python:**

```
vectorizer=TfidfVectorizer()
X = vectorizer.fit_transform(svm_title)
classifier=svm.SVC(kernel='linear', C=0.1, decision_function_shape='ovo')
classifier.fit(X,all_label)
svm_predict = classifier.predict(X)
print(classification_report(all_label,svm_predict,
target_names=list(data_dict.keys())))
```

- **Cài đặt mô hình PHOBERT trên ngôn ngữ Python:**

```
Config=AutoConfig.from_pretrained('vinai/phobertbase',output_hidden_states
=True)
phobert_model =TFAutoModel.from_pretrained ('vinai/phobertbase', config
=config)
tokenize =AutoTokenizer.from_pretrained('vinai/phobert-base')
def infer_data (embed_data, target_model):
    label_predict = []
    for line in tqdm(embed_data):
        line = tf.convert_to_tensor(line)
        line = tf.reshape(line, (1,3072))
        predict = model.predict([line])
        predict = np.argmax(predict, axis=1)
        label_predict.append(predict[0])
    return label_predict
results = infer_data(phobert_embed, model)
print (classification_report (all_label, results, target_names=list (data_dict.
keys())))
```

- **Cài đặt mô hình BERT multilingual trên ngôn ngữ Python:**

```
multilingual_tokenizer=BertTokenizer.from_pretrained('bert-base-
multilingual-cased')
multilingual_model = TFBertModel.from_pretrained("bert-base-multilingual-
cased")
results = infer_data (multilingual_embed, model_multilangual)
print (classification_report (all_label, results, target_names=list (data_dict.
keys())))
```

3.5 Các mô hình thực nghiệm

Luận văn đã thực hiện 2 loại gán nhãn cho tiêu đề truyền hình với việc sử dụng 3 dạng mô hình khác nhau để so sánh là: SVM, BERT multi language và PHOBERT.

❖ Mô hình SVM

Mô hình SVM luận văn thực nghiệm sử dụng pipeline để thực hiện các bước theo trình tự với một đối tượng, dùng TfidfVectorizer để thay đổi vector văn bản được tạo bởi bộ vector đếm và dùng hỗ trợ máy vector LinearSVC.

❖ Mô hình BERT multilingual

BERT multilingual là một mô hình của google BERT đa ngôn ngữ. Mô hình được đào tạo trước trên 100 ngôn ngữ hàng đầu có Wikipedia lớn nhất bằng cách sử dụng với mục tiêu tạo ra mô hình ngôn ngữ bị che (masked language modeling - MLM). Mô hình được phân biệt chữ hoa và chữ thường.

Luận văn được sử dụng mô hình huấn luyện cho trước bert-base-multilingual-cased. Trong mô hình huấn luyện luận văn sử dụng ClassificationModel của simpleTransformer để tạo ra mô hình huấn luyện. Luận văn thực hiện huấn luyện với số lượng train epochs là 1.

❖ Mô hình PHOBERT

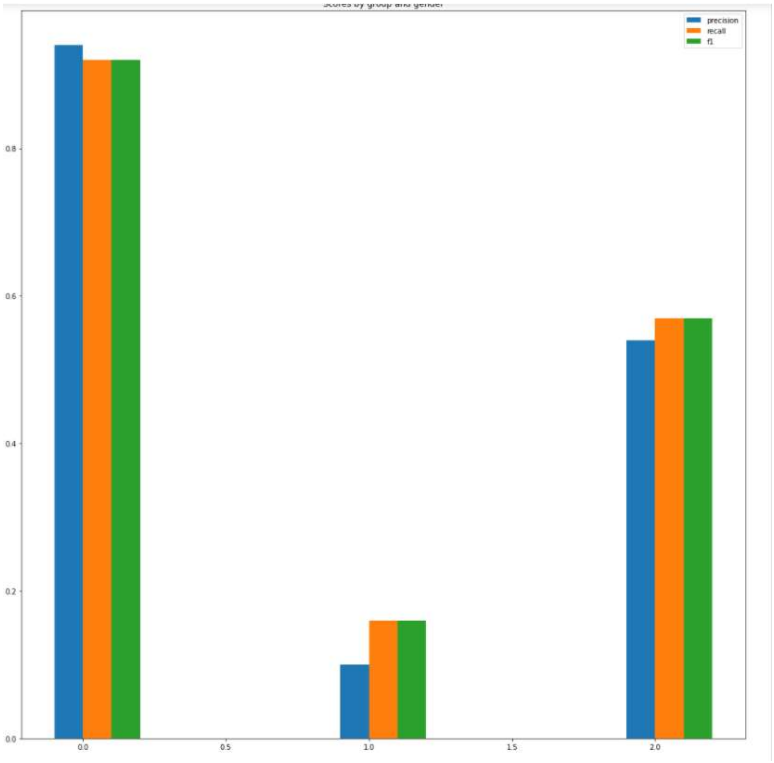
PHOBERT là mô hình huấn luyện trước, đặc biệt chỉ huấn luyện dành riêng cho tiếng Việt. PHOBERT huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa.

Tương tự như BERT, PHOBERT cũng có hai phiên bản là PHOBERT base với 12 transformers block và PHOBERT large với 24 transformers block.

Xây dựng model huấn luyện PHOBERT có hai lựa chọn là Fairseq và Transformer. Luận văn lựa chọn thử nghiệm với Transformer và sử dụng BertForSequenceClassification để tạo model. Trong phân loại binary luận văn thực hiện huấn luyện với số lượng epochs là 1, batch_size là 1000.

3.6 Kết quả thực nghiệm

Luận văn tiến hành làm thực nghiệm theo từng nhãn. Kết quả thực nghiệm từng phương pháp khá khả quan. Dưới đây là bảng kết quả mô tả các mô hình thực nghiệm.



Hình 3.5: Biểu đồ kết quả thực nghiệm phân loại của 3 mô hình

Bảng 3.3: Kết quả thực nghiệm phân loại của 3 mô hình

Mô hình	PRECISION(%)	RECALL(%)	F1(%)
SVM	54	57	53
BERT multilingual	1	16	13
PHOBERT	94	92	92

Từ bảng kết quả nhận thấy với độ đo F1 mô hình PhoBert cho kết quả tốt nhất (92%), cao hơn mô hình BERT multilingual (13%) và cao hơn mô hình SVM (53%)

Mô hình PhoBert cho kết quả tốt nhất.

Kết quả chi tiết cho từng nhãn được trình bày ở dưới đây:

Bảng 3.4: Kết quả thực nghiệm phân loại sử dụng mô hình SVM

STT	Nhãn	Precision(%)	Recall(%)	F1(%)
1	Phim truyện	100	83	91
2	Ca nhạc	100	3	6
3	Kỹ năng sống	0	0	0
4	Giải trí	35	100	52
5	Thời sự	98	20	33
6	Thể thao	0	0	0
7	Trẻ em	0	0	0
8	Du lịch	100	56	72

Bảng 3.5: Kết quả thực nghiệm phân loại sử dụng mô hình BERT

STT	Nhãn	Precision(%)	Recall(%)	F1(%)
1	Phim truyện	73	40	52
2	Ca nhạc	4	35	7
3	Kỹ năng sống	27	9	13
4	Giải trí	29	20	24
5	Thời sự	54	3	6
6	Thể thao	3	19	5
7	Trẻ em	4	16	6
8	Du lịch	4	3	3

Bảng 3.6: Kết quả thực nghiệm phân loại sử dụng mô hình PHOBERT

STT	Nhãn	Precision(%)	Recall(%)	F1(%)
1	Phim truyện	99	98	98
2	Ca nhạc	100	94	97
3	Kỹ năng sống	93	83	88
4	Giải trí	87	91	89
5	Thời sự	86	91	89
6	Thể thao	100	90	95
7	Trẻ em	100	89	94
8	Du lịch	98	92	95

Từ các bảng kết quả trên nhận thấy:

Kết quả phân loại nhị phân theo từng nhãn của các mô hình khá trên lệch. Các nhãn được phân loại theo mô hình PhoBert điều đạt kết quả khá tốt, đều trên 85%.

CHƯƠNG 4: ĐÁNH GIÁ KẾT QUẢ THỬ NGHIỆM

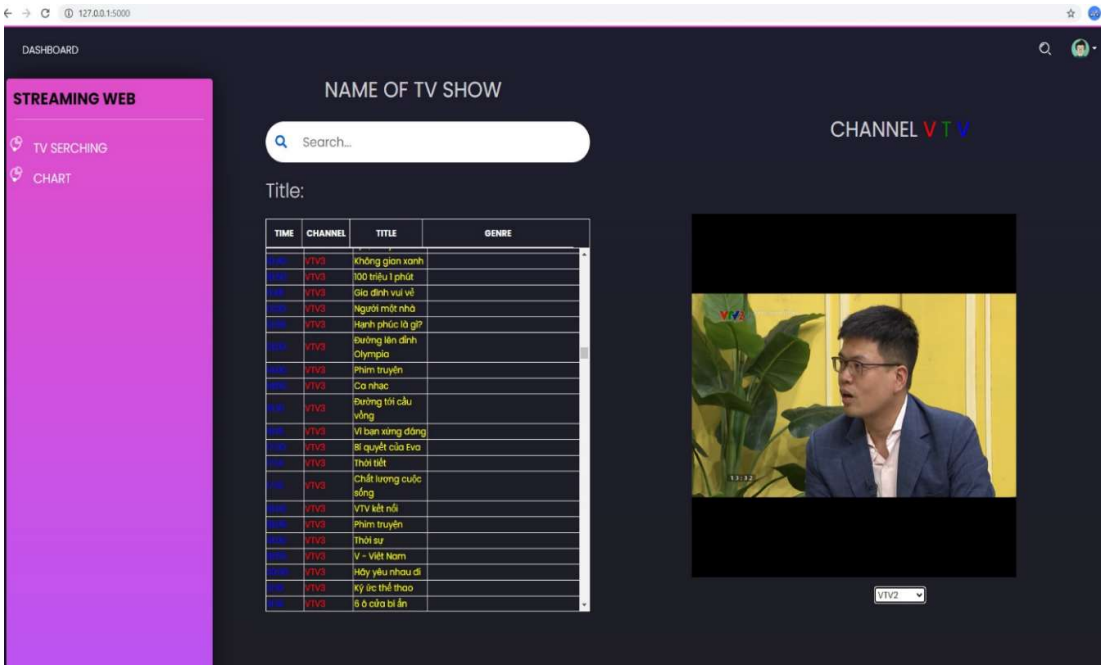
Chương này mô tả chi tiết về việc thử nghiệm cải tiến mô hình OTT trong lĩnh vực phân loại nội dung chương trình phát.

4.1 Mô tả kết quả phân loại chương trình

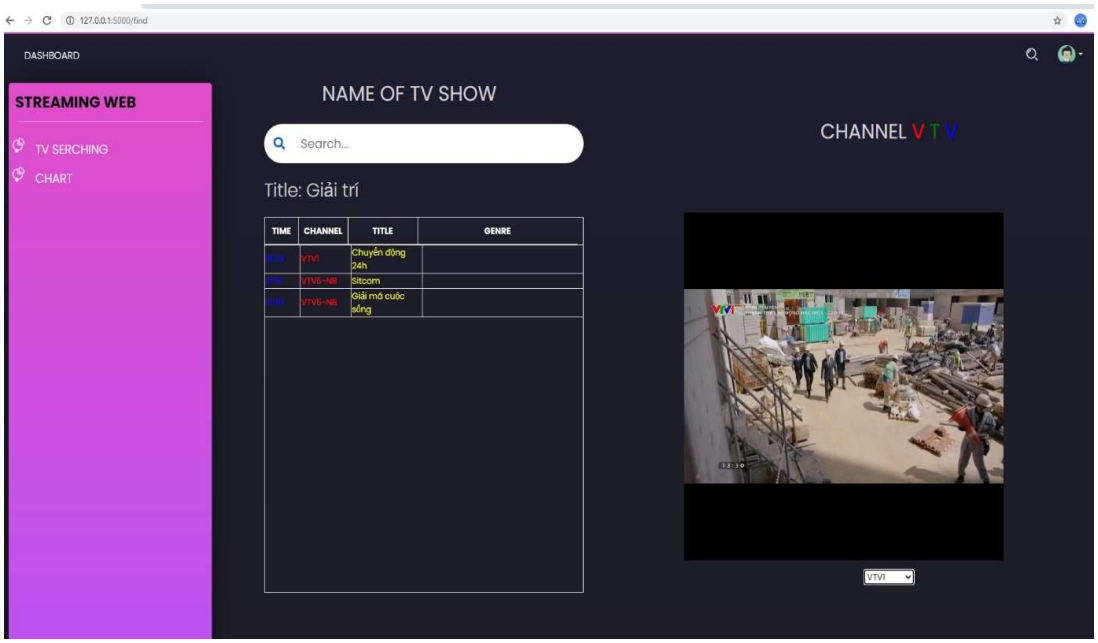
❖ Giao diện chương trình:

Hình 4.1 là giao diện hiển thị danh sách các kênh truyền hình được trình chiếu trong ngày. Người dùng nhập vào ô tìm kiếm để tìm danh sách các kênh đang được trình chiếu trong khung giờ của lịch phát sóng VTV.

Ví dụ: Người dùng nhập "Thể thao 24 giờ" thì danh sách theo chủ đề thể thao sẽ hiện thị lên theo các kênh của VTV. Người dùng chọn kênh theo nôi dụng mà họ yêu thích hoặc tìm các khung giờ mà họ có thể xem.

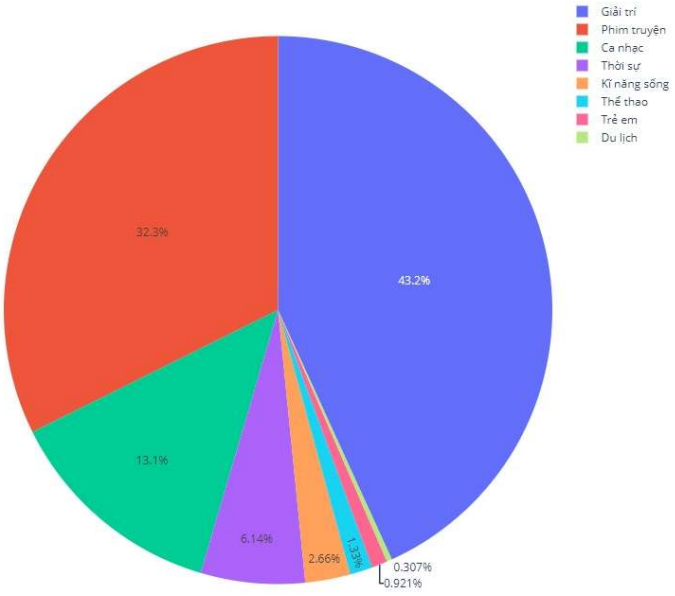


Hình 4.1: Giao diện danh sách lịch phát sóng VTV



Hình 4.2: Giao diện tìm kiếm nội dung theo sở thích của người dùng

❖ Biểu đồ:



Hình 4.3: Giao diện biểu đồ theo từng nhãn của chương trình

4.2 Kết luận

Đề tài phân loại chương trình truyền hình Internet theo nội dung là một trong những phương pháp giúp cho người xem tiếp cận với nội dung truyền hình trong khoảng thời gian ngắn nhất.

Với tính năng tương thích chương trình, lịch xem truyền hình quen thuộc được phân tích lại. Kênh truyền hình không còn giữ vai trò quan trọng trong việc tìm kiếm nội dung cần xem. Từ khóa tìm kiếm ở đây là sự phân loại thực sự của chương trình phát sóng. Việc tìm kiếm từ khóa có trong tên chương trình, việc sắp xếp các chương trình theo chủ đề của nội dung sẽ cải tạo hoàn toàn giao diện của lịch phát sóng cung cấp trong đề tài.

Đề tài đã nghiên cứu khái quát về các tựa đề của chương trình Truyền hình và tự động phân tích ngữ nghĩa theo phương pháp xử lý ngôn ngữ tự nhiên để phân lớp thành nhiều tuần dữ liệu của các chương trình truyền hình, tập hợp các từ khóa thường gặp cho mỗi chủ đề để gợi ý trước cho người xem. Tên tựa đề thường là 1 mệnh đề không trọn vẹn (không đủ thành 1 câu trọn nghĩa), việc phân tích bằng học sâu sẽ không thể cho kết quả tốt. Tựa đề là tiếng Việt Nam, các kiểu chơi chữ trong dùng từ để tiêu đề thêm súc tích, v.v… cũng gây ảnh hưởng không nhỏ đến kết quả dự đoán của hệ thống.

Việc mất cân bằng trong chủ đề phát sóng của truyền hình (trẻ em và du lịch chỉ chiếm phần nhỏ trong toàn chương trình truyền hình) cũng tác động đến sự sai lệch trong việc dự đoán chủ đề. Việc sử dụng mô hình PhoBert được huấn luyện trước với nhiều từ việt ngữ hơn, việc tự động tìm kiếm qua các nguồn tìm kiếm trên Internet để hiểu thêm môi trường cho những tựa đề của chương trình, làm phong phú thêm về số lượng và chất lượng của tập tin được học thêm cho mô hình huấn luyện là những hướng đi khả thi và hứa hẹn sẽ cải thiện tốt hơn việc nhận biết / phân loạt tên chương trình một cách nhanh chóng và chính xác.

4.3 Kiến nghị hướng nghiên cứu tiếp theo

Hướng nghiên cứu tiếp theo là tiến hành cài đặt đánh giá phương pháp tự động để phân bổ nội dung theo kênh phát trong môi trường Internet thực tế.

Đề tài cần nghiên cứu phát triển thêm các giải pháp, thuật toán AI (Deep learning, Machine learning) để phân tích điều khiển dạng Text và giọng nói để phân loại đáp ứng nhu cầu thực tế, giúp cho người dùng có thể tìm kiếm nhanh nội dung chương trình truyền hình muốn xem, gợi ý cho người dùng các chủ đề theo sở thích, thói quen trong khoảng thời gian ngắn nhất và chính xác nhất.

Cải thiện giao diện để thân thiện hơn cho người dùng bằng tính năng tương thích theo chương trình phát sóng. Đề tài đã đề xuất một phương thức khác hẳn với truyền hình truyền thống để xem TV. Người xem sẽ xác định chủ đề mình muốn xem khi bật TV.

4.4 Các công trình bài báo nghiên cứu

[1]. Võ Quang Long, Nguyễn Ngọc Hùng Anh, TS.Trần Minh Sơn, PGS.TS.Trần Thu Hà. Phân Loại Tên Chương Trình Truyền Hình Theo Chủ Đề Phát Sóng Sử Dụng Mô Hình Xlnet, Trường đại học Sư phạm Kỹ thuật TP.HCM, năm 2021.

[2]. Nguyễn Ngọc Hùng Anh, Võ Quang Long, TS.Trần Minh Sơn, PGS.TS.Trần Thu Hà, Giới Thiệu Về Zabbix, Hệ Thống Giám Sát Thường Xuyên Tài Nguyên Của Máy Chủ, Trường đại học Sư phạm Kỹ thuật TP.HCM, năm 2021.

DANH MỤC TÀI LIỆU THAM KHẢO

[1]. Trang web chính thức của Thủ Đô Multimedia: http://thudomultimedia.vn/-truyen-hinh-ott-xu-huong-tat-yeu-cua-truyen-hinh-thoi-dai-moi/, truy cập ngày 20/04/2021

[2]. Trang web chính thức của WebRTC: https://webrtc.org, truy cập ngày 28/10/2021

[3]. Lịch phát sóng các đài truyền hình VN: https://lichphatsongtivi.com/, truy cập ngày 30/04/2021

[4]. Mô hình phân tích đoạn văn tiếng Việt: https://www.vinai.io/phobert-the-first-public-large-scale-language-models-for-vietnamese/, truy cập ngày 28/10/2021

[5]. Trang web chính thức của Điện Tử Ngày Nay: https://dientungaynay.vn/-tags/truyen-hinh-ott, truy cập ngày 25/04/2021

[6]. Trang web chính thức của VNPT: https://vnpt.com.vn/tu-van/truyen-hinh-ott-la-gi.html, truy cập ngày 20/09/2021

[7]. A. Punchihewa, *Tutorial on IPTV and its latest developments*, ICIAFS January 2011

[8]. Trang web chính thức của Wikipedia: https://vi.wikipedia.org/wiki /Truy%E1%BB%81n_h%C3%ACnh_giao_th%E1%BB%A9c_Internet truy cập ngày 10/06/2021

[9] Trang web chính thức của FPT: https://hcmfpt.vn/ott-la-gi-tai-sao-noi-ott-la-xu-huong-khong-the-tranh-khoi.html, truy cập ngày 28/04/2021

[10]. T. Ohanian, *Over-the-Top Considerations: Functionalities and Technologies* Cisco Systems, NAB 2014

[11]. C. Waldenor, Is OTT Disrupting Television? Master Thesis, Stockholm, June 7th 2013

[12]. Nguyễn Minh Thành, Phân loại văn bản, Đồ án môn học Xử lý ngôn ngữ tự nhiên, Đại học quốc gia Thành phố Hồ Chí Minh, 01/2011

[13]. Nguyễn Thị Hương Thảo, Phân lớp phân cấp Taxonomy văn bản Web và ứng dụng. Khóa luận tốt nghiệp đại học, Đại học Công nghệ, 2006

[14]. Mô hình Bert: https://phamdinhkhanh.github.io/2020/05/23/BERTModel.html, truy cập ngày 10/11/2021

[15]. Trang web chính thức Machinelearning: https://machinelearningcoban.com/2016/12/27/categories/, truy cập ngày 20/08/2021

[16]. Mô hình phân loại văn bản tiếng việt: https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/, truy cập ngày 22/11/2021

[17]. Trang web chính thức Machinelearning: https://machinelearningcoban.com/2017/01/01/kmeans/, truy cập ngày 25/08/2021

[18]. Mô hình thuật toán K-Means: http://bis.net.vn/forums/t/374.aspx, truy cập ngày 22/10/2021

[19]. Mô hình thuật toán K-Means: https://machinelearningcoban.com/2017/01/01-/kmeans/, truy cập ngày 20/08/2021

[20]. Mô hình thuật toán Bert: https://blog.vietnamlab.vn/gioi-thieu-bert-va-ung-dung-vao-bai-toan-phan-loai-van-ban/, truy cập ngày 15/08/2021

[21]. Mô hình thuật toán Bert: https://viblo.asia/p/hieu-hon-ve-bert-buoc-nhay-lon-cua-google-eW65GANOZDO, truy cập ngày 15/10/2021

[22]. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, Đại học Carnegie Mellon, Nhóm trí tuệ AI của Google.

[23]. Giới thiệu về chuyển đổi câu trong xử lý ngôn ngữ tự nhiên: https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/, truy cập ngày 28/10/2021

[24]. Mô hình thuật toán Bert: https://buiminhptit.github.io/2020/03/10/gi%E1%BA%-A3i-th%C3%ADch-m%C3%B4-h%C3%ACnh-transformer.html, truy cập ngày 20/10/2021

[25]. Lịch phát sóng VTV: https://vtv.vn/lich-phat-song.htm, truy cập ngày 20/09/2021

[26]. Trang web chính thức Machinelearning: https://machinelearningcoban.com/2017/08/31/evaluation/, truy cập ngày 28/11/2021