

# Bộ dữ liệu thông tin video trên nền tảng Youtube và xây dựng mô hình phân loại video

Ngô Phước Thịnh<sup>1[19520981]</sup> và Đặng Trần Anh Khoa<sup>2[19520629]</sup>

<sup>1, 2</sup> Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh,  
Thành phố Thủ Đức, 70000 Thành phố Hồ Chí Minh, Việt Nam

*Abstract*-Youtube hiện nay đang là nền tảng chia sẻ video lớn nhất trên thế giới với nội dung phong phú, đa dạng hướng đến nhiều đối tượng người dùng khác nhau. Youtube không chỉ là nền tảng để đăng tải các video giải trí, chia sẻ khoảnh khắc mà người đăng tải hiện nay còn có nhiều mục đích khác như tuyên truyền quảng bá, kiếm tiền, quảng cáo,... hoặc thậm chí có nhiều mục đích xấu khác như tuyên truyền, bịa đặt,... cần được ngăn chặn. Do đó, việc phân tích các video trên Youtube là một việc làm hết sức cần thiết để có thể nhanh chóng nắm bắt được xu hướng đang thịnh hành, phân loại các video, ngăn chặn các video tiêu cực ,... Vì vậy nhóm đã chọn chủ đề cho đồ án môn học là “Bộ dữ liệu thông tin các video trên nền tảng Youtube và xây dựng mô hình phân loại video” để giới thiệu một quy trình thu thập, tiền xử lý và dán nhãn để có được một bộ dữ liệu phù hợp với các tác vụ trên, đồng thời cũng thực hiện phân tích một vài khía cạnh của bộ dữ liệu, thử nghiệm thực tế trên các mô hình máy học.

*Keywords:* dataset, classification

## 1 Giới thiệu bộ dữ liệu

Bộ dữ liệu thông tin các video trên nền tảng Youtube bao gồm 9 thuộc tính với 10734 dòng dữ liệu.

9 thuộc tính của bộ dữ liệu bao gồm:

- **ID:** ID của video trên Youtube, gồm 11 kí tự base64
- **Title:** Tiêu đề của video trên Youtube
- **Description:** Phân mô tả của video
- **Channel ID:** ID của kênh đăng tải video, gồm 24 kí tự base64
- **Channel Subscriber count:** Số người đăng kí của kênh đăng tải video, làm tròn đến hàng nghìn
- **Channel view count:** Số lượng lượt xem trên kênh
- **View count:** Số lượt xem của video
- **Like count:** Số like trên kênh
- **Duration:** Thời lượng video

Bộ dữ liệu hướng đến các Video có ngôn ngữ tiếng Việt hoặc có nhiều lượt xem ở vùng Việt Nam, hoặc liên quan đến Việt Nam. Các video trong bộ dữ liệu thu thập được hướng đến 15 chủ đề được phân loại sẵn theo Youtube Category[1] bao gồm:

*Film & Animation*

*Autos & Vehicles*

*Music*

*Pet & Animal*

*Sport*

*Travel & Events*

*Gaming*

*People & Blogs*

*Comedy*

*Entertainment*

*News & Politics*

*Howto & Style*

*Education*

*Science & Technology*

*Nonprofits & Activism*

Đồng thời các chủ đề trên cũng sẽ là nhãn của các dòng dữ liệu. Mỗi điểm dữ liệu sẽ mang một nhãn ứng với 1 trong 15 chủ đề ở trên.

## 2 Phương pháp thu thập

Sau khi xác định được nội dung và đối tượng cần thu thập mà bộ dữ liệu hướng đến, nhóm tiếp tục tiến hành giai đoạn thu thập dữ liệu.

Dữ liệu được nhóm thu thập thông qua API, ở đây nhóm sử dụng Youtube Data API [2] để lấy thông tin thông qua Script được viết trên Google App Script (Extension được tích hợp sẵn trong Google Sheet – tiện ích trang tính của Google) và lưu lại dữ liệu thô trên các file Google Sheet. Bằng cách đưa ra các từ khóa liên quan đến chủ đề, Youtube Data API sẽ trả về các video liên quan đến các từ khóa.

## 3 Tiền xử lí

Sau khi thu thập được dữ liệu thô, chúng ta cần phải thực hiện tiền xử lí để dữ liệu về đúng định dạng và đủ tiêu chuẩn một bộ dữ liệu sạch. Sau khi thu thập, nhiều dòng dữ liệu bị rỗng và còn có thuộc tính chưa đưa về đúng định dạng mong muốn, vì vậy ở bước này ta sẽ thực hiện xử lí các giá trị này.

Trước tiên, ta cần gộp 15 file ứng với 15 chủ đề lại thành một file. Vì sau khi thu thập bằng cách đưa ra từ khóa, ứng với mỗi chủ đề sẽ có 1 file raw data riêng vì vậy ta cần ghép các file này lại.

Tiếp theo cần loại bỏ các dòng dữ liệu bị trống cũng như các dòng dữ liệu bị trùng thông qua VideoID để đảm bảo các video không bị xuất hiện trùng lặp.

Sau khi xử lí các dòng dữ liệu trùng, ta tiếp tục xử lí về kiểu dữ liệu, đảm bảo một số thuộc tính ở đúng định dạng numeric như Channel view count, Channel Subscriber count, Like count. Thứ 2 là ở định dạng của cột Duration, định dạng dữ liệu của cột này được API trả về theo chuẩn ISO-8601[3] có dạng PTHMS, chúng ta sẽ đưa về định dạng HH:MM:SS.

## 4 Gán nhãn

### 4.1 Qui trình gán nhãn

Sau quá trình tiền xử lí, ta cần gán nhãn cho bộ dữ liệu. Nhóm thực hiện việc dán nhãn qua 2 giai đoạn: dán nhãn trên mẫu 299 dòng để đánh giá độ đồng thuận và dán nhãn diện rộng. Đầu tiên, nhóm sẽ xây dựng bộ hướng dẫn dán nhãn - Guideline. Guideline sẽ trình bày các quy tắc để dán được nhãn cho bộ dữ liệu.

**Định nghĩa các nhãn.** Các nhãn của dữ liệu được định nghĩa như sau:

**Table 1:** Định nghĩa các nhãn trong bộ dữ liệu

Nhãn	Định nghĩa
<b>Film &amp; Animation</b>	Gồm 2 lĩnh vực là Phim ảnh và Hoạt hình. Phim ảnh bao gồm phim, trailer, nhạc phim, các cảnh quay trong phim, hậu trường phim, và những video có nội dung liên quan. Hoạt hình bao gồm các video hoạt hình ngắn, chuyện ngắn, video kể chuyện và những video hoạt ảnh khác liên quan. Danh mục này thường được sử dụng bởi những thương hiệu lớn và các nhà sản xuất, liên quan trực tiếp đến ngành công nghiệp điện ảnh và hoạt hình.
<b>Autos &amp; Vehicles</b>	Danh mục Ô tô và xe gồm những video liên quan đến ô tô, công nghệ ô tô, các loại xe,... Ngoài ra còn có các video dạng review, mẹo, giải pháp, ra mắt, và các dạng video khác có liên quan đến chủ đề.
<b>Music</b>	Bao gồm tất cả các thể loại nhạc, bài hát, và những video liên quan. Có thể bao gồm nhạc cụ và các hướng dẫn thanh nhạc. Đây là danh mục phổ biến nhất, có số lượng lớn người đăng kí, theo dõi.
<b>Pets &amp; Animals</b>	Bao gồm các video về thú cưng, động vật, thức ăn cho động vật, mẹo chăm sóc, cũng như các sản phẩm liên quan đến động vật, các video vui nhộn, và các video khác liên quan.
<b>Sports</b>	Danh mục thể thao bao gồm các video về thể thao, thiết bị thể thao, các mẹo, các khoảnh khắc thú vị, các thống kê, hoặc các hướng dẫn tập thể thao, hoặc các hoạt động khác liên quan.
<b>Travel &amp; Events</b>	Danh mục bao gồm các mẹo vật du lịch, địa điểm du lịch, các nơi nghỉ dưỡng, sự kiện, tổ chức các sự kiện, và các video khác có liên quan.
<b>Gaming</b>	Danh mục trò chơi bao gồm các thông tin về game, các mẹo, các đánh giá, và các video chơi game, cũng như các video khác liên quan.
<b>People &amp; Blogs</b>	Gồm các video cho con người, lối sống, người nổi tiếng, các vấn đề về con người, các blog phổ biến, hoặc các trang thông tin, đánh giá liên quan đến con người và blogs.
<b>Comedy</b>	Gồm các video hài kịch, có thể bao gồm những bài nói, câu chuyện ngắn, phim hoạt hình, hoặc các video vui nhộn. Hay nói cách khác, danh mục này chứa các video gây cười.
<b>Entertainment</b>	Giải trí bao gồm nhiều chủ đề, từ nhảy, kịch, chuyện kể,... Kiểm tra các danh mục có liên quan (phim ảnh, âm nhạc, hài kịch) trước khi xếp vào danh mục giải trí.
<b>News &amp; Politics</b>	Danh mục này bao gồm các video tin tức ngắn, các video thời sự, các tin chính trị, và các video liên quan.
<b>Howto &amp; Style</b>	Gồm các video về cách làm, và các video liên quan đến thời trang.

<b>Education</b>	Danh mục giáo dục bao gồm các hướng dẫn, các chỉ dẫn mang kiến thức, các phương pháp học, lớp học, và các video giàu thông tin khác.
<b>Science &amp; Technology</b>	Bao gồm các video thuộc lĩnh vực khoa học, công nghệ. Ví dụ như các cải tiến, các sự thật, các công nghệ tương lai, công nghệ di động, máy tính,...
<b>Nonprofits &amp; Activism</b>	Gồm các hoạt động phi lợi nhuận và các chiến dịch quảng bá.

**Quy tắc xác định nhãn.** Sau khi định nghĩa các nhãn, ta sẽ định nghĩa các quy tắc để xác định được nhãn.

*Trường hợp 1:* Thông qua tiêu đề có thể xác định được thể loại của video để dàng thông qua từ khóa và ngữ cảnh.

Ví dụ:

- Tiêu đề “Bản tin Dự báo Thời tiết đêm 10, ngày 11/5/2022” sẽ được phân loại vào nhãn **News & Politics**.
- Tiêu đề “CẬP TRƯỢT BĂNG NGHỆ THUẬT NHẬT BẢN TẠI THỂ VẬN HỘI MÙA ĐÔNG BẮC KINH 2022” sẽ được phân loại vào nhãn **Sports**.
- “Lập Sườn Cao Bằng - Bản tin đặc sản vùng miền của Đài Truyền Hình Nhân Dân 098.339.3412” được phân loại vào **Travel & Events**.

*Trường hợp 2:* Trong trường hợp thông qua tiêu đề chưa thể xác định được nhãn của video thì ta sẽ dựa vào phần mô tả.

Ví dụ:

- Với tiêu đề “Cách Sử Dụng Giới Từ Cho Phương Tiện Giao Thông” có thể được phân loại vào **Auto & Vehicles** hoặc **Education**.  
Dựa vào mô tả: “Chào mừng các bạn đến với chương trình học tiếng Anh Online cùng với trường Anh ngữ CIP. Mình là Thảo. Hôm nay mình sẽ ...”, video có thể được phân loại vào lớp **Education**.

*Trường hợp 3:* Trong trường hợp tiêu đề lẫn mô tả của video mang 2 chủ đề khác nhau sẽ chọn chủ đề chính và quan trọng nhất của video.

Ví dụ:

- Với tiêu đề “Phương tiện giao thông/BÉ NGHE NHÌN VÀ ĐỌC TÊN CÁC LOẠI XE QUEN THUỘC TRÊN ĐƯỜNG PHỐ/nhạc vui nhộn.” Video trên thuộc về 3 chủ đề là **Auto & Vehicle**, **Education**, **Music**, nhưng mục đích của video hướng đến giáo dục trẻ em nên nhãn được chọn là **Education**.

*Trường hợp 4:* Thông qua tiêu đề và mô tả vẫn chưa xác định được thì sẽ sử dụng VideoId để lấy đường dẫn đến video trên Youtube, và người gán nhãn sẽ xem video sau đó quyết định nhãn.

**Dán nhãn.** Sau khi đọc kỹ định nghĩa của các nhãn và hướng dẫn dán nhãn, 2 người sẽ tiến hành dán nhãn độc lập với nhau. Khi 2 người dán nhãn hoàn tất việc dán nhãn, ta sẽ tiến hành đánh giá độ đồng thuận.

#### 4.2 Đánh giá độ đồng thuận

Sau khi gán nhãn, ta tính được độ đồng thuận như sau:

Xác suất đồng thuận giữa các nhãn:

$$\mathbf{Pr(a)} = 0.6454849498327759$$

Xác suất độ đồng thuận của 2 người trên từng nhãn:

**Table 2:** Độ đồng thuận của 2 người dán nhãn trên từng nhãn

Nhãn	Độ đồng thuận
Autos & Vehicles	0.0005480923032180848
Comedy	0.0013534524222324137
Education	0.004037986152280176
Entertainment	0.00071587566134607
Film and Animation	0.0025167503719197774
Gaming	0.0009060301338911198
Howto & Style	0.00286350264538428
Music	0.00591715976331361
News & Politics	0.005413809688929655
Nonprofits & Activism	1.1185557208532343e-05 ~ 0

Xác suất giả định của 2 người dán nhãn trên 15 nhãn:

$$\mathbf{Pr(e)} = 0.03318754823771546$$

Độ đồng thuận của 2 người dán nhãn:

$$\mathbf{K} = 0.6333155934007451$$


Theo thang đo độ đồng thuận của Cohen & Kappa[4], mức đồng thuận này ở mức “Good” (0.61-0.80), nhóm có thể tiếp tục sử dụng phương pháp này để thực hiện dán nhãn diện rộng trên toàn bộ dữ liệu.

## 5 Phân tích sơ bộ dữ liệu

Sau khi hoàn tất dán nhãn 600 dòng của bộ dữ liệu, nhóm đã tiến hành phân tích và khám phá bộ dữ liệu. Bộ dữ liệu bao gồm 9 đặc trưng bao gồm: 'channelId', 'videoId', 'title', 'Description', 'ChannelViewCount', 'subscriberCount', 'viewCount', 'likeCount', 'Duration'.

Tiêu đề của 10 video có lượt xem cao nhất bao gồm:

- Giấc Mơ Làm Mẹ - Câu Chuyện Cảm Động Về Tình Mẫu Tử Thiêng Liêng | Phim Hoạt Hình Việt Nam”
- BUỒN LÀM CHI EM ƠI ► ĐOẠN TUYỆT ► LK Nhạc Vàng Bolero Toàn Bài Hay KHÔNG QUẢNG CÁO HAY NỨC LÒNG
- Con Kiến Con, Chú Vịt Con - Nhạc Thiếu Nhi Vui Nhộn
- Eena Meena Deeka Superclip 7 - Cuộc Rượt Đuổi Của Cáo Và Gà - Funny Cartoon

- LA LA SCHOOL | CHỊ ĐẠI LYKIO | Season 1 : Học Viện Siêu Sao (Phim Ca Nhạc Học Đường 2017)
-  HỌC SINH CÁ BIỆT | Biệt Đội Biết Tuốt | Phim Học Đường Vui Nhộn MIU MIU SCHOOL
- Con Gà Trống Thối Kèn, Con Cò Bé Bé - Ca nhạc thiếu nhi vui nhộn cho bé yêu
- Phim hoạt hình - Ở BẮN NHẤT XÓM - Truyện Cổ tích - Quà tặng cuộc sống - Nghệ thuật sống
- Nghe Nhạc Thư Giãn Rumba Không Lời Ngắm Cảnh Thiên Nhiên Hùng Vĩ, Giảm Stress Quên Hết Mọi
- Bài Hát Làm Cả Thế Giới Khóc Nghiêng Ngã - Nhạc Chế Kiếp Lục Bình Trời Đã Đứng Trái Tim Người Nghe

Trong 10 video có lượt xem cao nhất nhóm thu thập được thì có đến 5 video là chủ đề Music, 4 trong 5 video đó hướng đến chủ đề dành cho trẻ em.

Khi thống kê 10 kênh có nhiều lượt đăng kí nhất gồm có:

<i>BLACKPINK</i>	<i>FAPTV</i>
<i>Masha and The Bear</i>	<i>GMA News</i>
<i>Vlad và Nikita</i>	<i>Kênh Thiếu Nhi - BHMEDIA</i>
<i>Triggered Insaan</i>	<i>Cris Devil Gamer</i>
<i>POPS Kids</i>	<i>NTN</i>
<i>Mobile Legends: Bang Bang Official</i>	<i>Vie Channel - HTV2</i>
<i>Like Nastya VNM</i>	<i>Soothing Relaxation</i>
<i>MNCTV OFFICIAL</i>	

Trong số tất cả các kênh có lượt theo dõi nhiều nhất thì 5/15 kênh làm về chủ đề hướng đến trẻ em (Masha and The Bear, Vlad và Nikita, POPS Kids, Like Nastya, Kênh thiếu nhi-BHMEDIA), các kênh còn lại hướng đến nhiều chủ đề Gaming, Music, Entertainment.

Trong số 15 kênh có lượt xem cao nhất bao gồm:

<i>BLACKPINK</i>	<i>VNM</i>
<i>Masha and The Bear</i>	<i>MNCTV OFFICIAL</i>
<i>Vlad và Nikita</i>	<i>BabyBus - Nhạc thiếu nhi</i>
<i>GMA News</i>	<i>Tony TV</i>
<i>Kênh Thiếu Nhi - BHMEDIA</i>	<i>Thơ Nguyễn</i>
<i>POPS Kids</i>	<i>SlenderMan™</i>
<i>POPS MUSIC</i>	<i>Thuy Nga</i>
<i>Like Nastya</i>	<i>KN Channel</i>

Trong số các kênh có lượt xem cao thì đã có sự thay đổi so với các kênh có lượt đăng kí nhiều nhất, tuy nhiên các kênh làm nội dung trẻ em vẫn chiếm một số một số lượng lớn.

Qua phân tích về video có lượt xem cao cùng với các kênh có lượt xem và lượt đăng kí cao nhất thì ta thấy các nội dung về trẻ em được quan tâm nhiều và có lượng người xem lớn, điều này có thể đến từ đặc điểm người xem hoặc xu hướng của phần lớn người dùng hiện nay thường xuyên chọn các video giải trí như hoạt hình.

## 6 Mô hình máy học

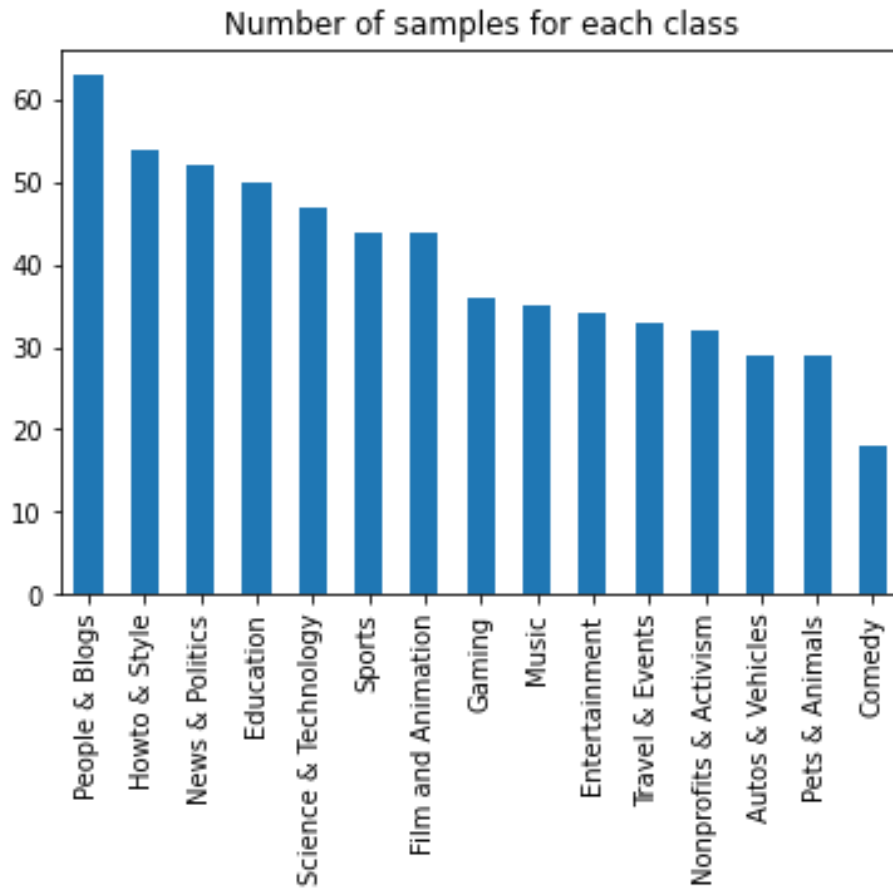
### 6.1 Xử lý dữ liệu và mã hóa

Sau khi hoàn tất việc dán nhãn, chúng ta có được một bộ dữ liệu hoàn chỉnh. Ở đây nhóm đã gán nhãn trên 600 điểm dữ liệu được chọn ngẫu nhiên. Tiếp theo, ta sẽ thực hiện huấn luyện một số mô hình máy học. Nhóm chọn 2 mô hình phân lớp là Naïve Bayes và SVM (Support Vector Machine).

Trước hết, ta cần mã hóa các nhãn thành dạng numeric để có thể đưa vào các mô hình máy học. Đây là bước quan trọng trước khi đưa dữ liệu vào các mô hình học giám sát. 15 nhãn trên bộ dữ liệu của chúng ta sẽ được mã hóa từ 0 đến 14.

**Table 3:** Các nhãn được mã hóa và giá trị tương ứng

Mã hóa	Nhãn
0	Autos & Vehicles
1	Comedy
2	Education
3	Entertainment
4	Film and Animation
5	Gaming
6	Howto & Style
7	Music
8	News & Politics
9	Nonprofits & Activism
10	People & Blogs
11	Pets & Animals
12	Science & Technology
13	Sports
14	Travel & Events



**Figure 1:** Số lượng các điểm dữ liệu theo từng nhãn

Tiếp theo ta thực hiện trích xuất đặc trưng từ bộ dữ liệu để phù hợp với mô hình mà chúng ta sử dụng, ở đây nhóm sẽ giữ lại 2 thuộc tính **Title** và **Description**. Hai thuộc tính sẽ được kết hợp lại theo phép cộng chuỗi. Sau đó xử lý loại bỏ các kí tự đặc biệt xuất hiện trong các tiêu đề như “^”, “[”, ... Thuộc tính kết hợp này sẽ được vector hóa bằng phương pháp TfidfVectorizer.

Nhóm tiến hành so sánh hiệu suất của mô hình trên từng trường hợp ngram\_range (giới hạn độ dài từ được phép kết hợp khi thực hiện vector hóa) để có thể chọn được mô hình phù hợp.

Bộ dữ liệu được chia thành 2 tập: tập huấn luyện và tập kiểm thử theo tỉ lệ 7:3

## 6.2 Naive Bayes:

Với mô hình Naïve Bayes, mô hình được nhóm triển khai là MultiNomialNB - một thuật toán dựa trên định lí Naïve Bayes[5].

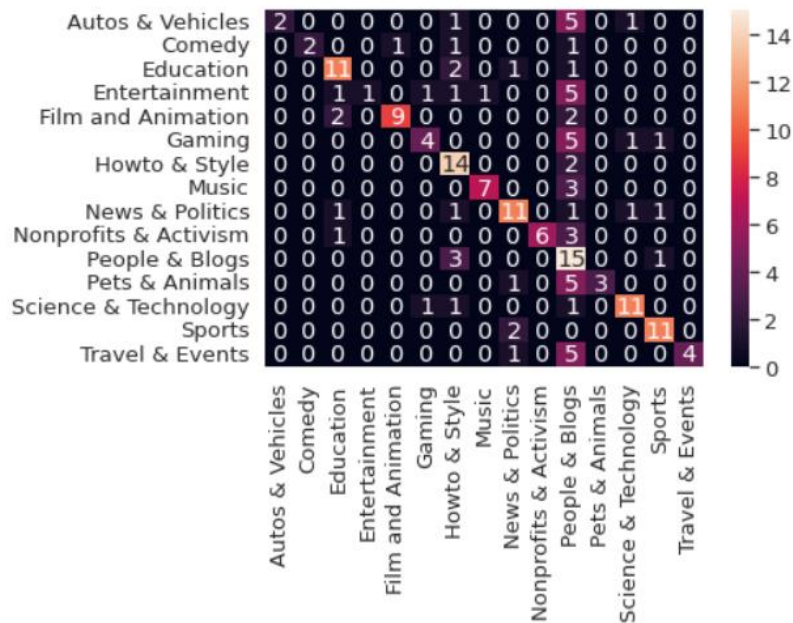


Nhóm thực hiện huấn luyện trên 3 Ngram-range lần lượt là (1,2), (1,3), (2,3). Kết quả thu được như sau:

**Table 4:** Kết quả mô hình Naïve Bayes dự đoán thông qua các độ đo

Ngram_range	F1 macro	Accuracy
(1-2)	59.1647036495096	61.11111111111111
(1-3)	59.34977165117340	60.55555555555550
(2-3)	48.1640132502685	51.66666666666660

Ta thấy với ngram\_range=(1,2) thì độ đo Accuracy đạt cao nhất và F1-Score không có khác biệt quá lớn với trường hợp ngram\_range=(1,3) nên nhóm sẽ chọn mô hình sử dụng ngram\_range=(1-2) cho các bước phân tích tiếp theo. Trực quan hóa dự đoán của mô hình trên heatmap:



**Figure 2:** Kết quả dự đoán của mô hình Naïve Bayes thể hiện qua heatmap

Kết quả phân loại:

**Table 5:** Kết quả phân loại mô hình Naive Bayes

	precision	recall	f1-score	support
0	1.00	0.22	0.36	9
1	1.00	0.40	0.57	5
2	0.69	0.73	0.71	15
3	1.00	0.10	0.18	10

4	0.90	0.69	0.78	13
5	0.67	0.36	0.47	11
6	0.58	0.88	0.70	16
7	0.88	0.70	0.78	10
8	0.69	0.69	0.69	16
9	1.00	0.60	0.75	10
10	0.28	0.79	0.41	19
11	1.00	0.33	0.50	9
12	0.79	0.79	0.79	14
13	0.79	0.85	0.81	13
14	1.00	0.40	0.57	10
<b>accuracy</b>			0.62	180
<b>macro avg</b>	0.82	0.57	0.61	180
<b>weighted avg</b>	0.77	0.62	0.62	180

Đánh giá mô hình trên độ đo F1 macro và độ chính xác:

**F1 macro:** 60.51967880686191

**Accuracy:** 61.66666666666667

Mô hình thực hiện dự đoán chính xác nhất trên nhãn số 13(Sports) và thấp nhất trên nhãn số 3( Entertainment).

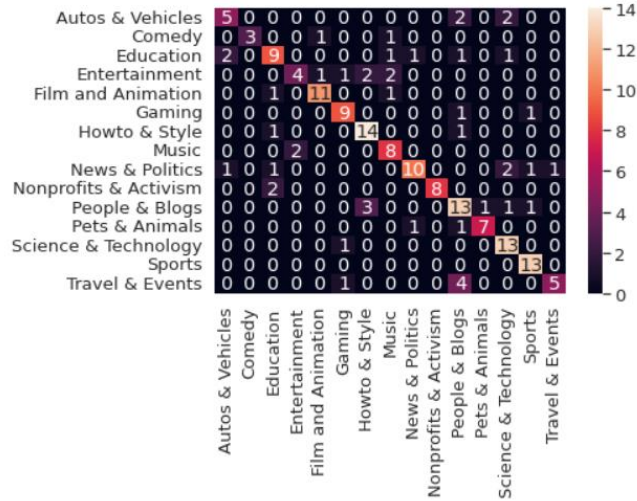
### 6.3 Support Vector Machine

Với mô hình SVM, mô hình được triển khai với một kernel của SVM tương tự với SVC là LinearSVC[6]. Bộ từ điển được xây dựng với 3 mức ngram\_range là (1,2), (1,3) và (2,3), độ lớn của từ điển là 20000. Sau khi thực hiện thì kết quả đạt được lần lượt là:

**Table 6:** Phân tích các ngram\_range

ngram_range	F1 macro	Accuracy
1-2)	72.58175889339320	72.77777777777777
(1-3)	72.92347874008514	73.33333333333333
(2-3)	66.44532415946244	66.11111111111111

Ta thấy với ngram\_range=(1,3) đạt được cao nhất ở cả độ đo Accuracy và F1 macro nên sẽ chọn ngram\_range=(1,3) để tiếp tục thực hiện phân tích. Trực quan hóa dự đoán của mô hình trên heatmap:



**Figure 3:** Kết quả dự đoán của mô hình SVM thể hiện qua heatmap

Kết quả phân loại:

**Table 7:** Kết quả phân loại của mô hình SVM

	precision	recall	f1-score	support
0	0.62	0.56	0.59	9
1	1.00	0.60	0.75	5
2	0.64	0.60	0.62	15
3	0.67	0.40	0.50	10
4	0.85	0.85	0.85	13
5	0.75	0.82	0.78	11
6	0.74	0.88	0.80	16
7	0.62	0.80	0.70	10
8	0.83	0.62	0.71	16
9	1.00	0.80	0.89	10
10	0.57	0.68	0.62	19
11	0.88	0.78	0.82	9
12	0.68	0.93	0.79	14
13	0.81	1.00	0.90	13
14	0.83	0.50	0.62	10
accuracy			0.73	180
macro avg	0.77	0.72	0.73	180
weighted avg	0.75	0.73	0.73	180

Đánh giá mô hình trên độ đo F1 macro và độ chính xác:

**F1 macro:** 72.92347874008514

**Accuracy:** 73.33333333333333

Ta thấy rằng với độ đo F1-score thì dự đoán ở trên nhãn số 13 (Sports) đạt cao nhất và nhãn thấp nhất là 3 (Entertainment).

## 7 Tổng kết và phương hướng tương lai

Sau khi thực hiện, cơ bản nhóm đã xây dựng được một bộ dữ liệu Thông tin các video trên Youtube từ việc thu thập, tiền xử lý, dán nhãn đến áp dụng vào bài toán thực tế là phân loại video dựa trên tiêu đề. Tuy nhiên vẫn còn nhiều thiếu sót trong bộ dữ liệu mà nhóm đã xây dựng được như: số lượng điểm dữ liệu còn quá thấp so với tổng thể video trên Youtube; một số nhãn chưa có sự phân định rõ ràng khi áp dụng từ Youtube Category khi áp dụng cho các nội dung tại Việt Nam (ví dụ như các lớp Entertainment, Music, Comedy). Bộ dữ liệu chỉ có ý nghĩa tức thời vì thông tin trên Youtube thay đổi liên tục theo thời gian nên dữ liệu nhanh chóng bị lỗi thời.

Trong tương lai, nhóm sẽ mở rộng bộ dữ liệu về cả số lượng điểm dữ liệu lẫn số thuộc tính để có thể đa dạng bộ dữ liệu hơn. Đồng thời xây dựng một hệ thống tự động thu thập định kì thông tin từ Youtube để dữ liệu được sát với thực tế nhất bởi thông tin ở Youtube thay đổi liên tục theo thời gian. Các nhãn có thể được thay đổi để có thể sát với tình hình thực tế và các xu hướng của hiện tại.

Mặc dù độ đồng thuận đang ở mức chấp nhận được và việc gán nhãn khá hiệu quả. Tuy nhiên nhóm cần phải cải thiện guideline để có thể đạt được độ đồng thuận tốt hơn và việc gán nhãn được chính xác hơn, tránh các trường hợp gán nhãn nhầm lẫn cũng như mô tả chính xác hơn về từng nhãn.

## References

1. How to Choose a Category on Youtube. Accessed: June 21, 2022. [Online] Available: <https://influenceonyoutube.com/create-videos-for-youtube-start-now/how-to-choose-a-category-on-youtube/>
2. Add YouTube functionality to your app, Accessed: June 21, 2022. [Online] Available: <https://developers.google.com/youtube/v3>
3. ISO 8601 DATE TIME FORMAT. Accessd: June 21, 2022. [Online] Available: <https://www.iso.org/iso-8601-date-and-time-format.html>
4. Calculating Kappa. Accessed: June 21, 2022. [Online] Available: [https://elentra.healthsci.queensu.ca/assets/modules/reproducibility/calculating\\_kappa.html](https://elentra.healthsci.queensu.ca/assets/modules/reproducibility/calculating_kappa.html)
5. Sklearn.naive\_bayes.MultinomialNB. Accessed: June 21, 2022. [Online] Available: [https://scikit-learn.org/stable/modules/generated/sklearn.naive\\_bayes.MultinomialNB.html](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html)
6. Sklearn.svm.LinearSVC. Access: June 21, 2022. [Online] Available: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>