ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN, ĐHQG-HCM

KHOA KHOA HỌC MÁY TÍNH

ĐỒ ÁN MÔN HỌC

CS221 - XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Đề tài: MULTI-LABEL MOVIE GENRES CLASSIFICATION FROM ORIGINAL OVERVIEW

Giảng viên hướng dẫn	**:**	**TS. Nguyễn Trọng Chỉnh**
Lớp	:	CS221.P22
Sinh viên thực hiện 1	:	Phùng Minh Chí
Mã số sinh viên 1	:	23520179
Sinh viên thực hiện 2	:	Nguyễn Hữu Minh Chiến
Mã số sinh viên 2	:	23520183
Sinh viên thực hiện 3	:	Lê Ngọc Phương Thảo
Mã số sinh viên 3	:	23521467

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

LỜI CẢM ƠN

Để hoàn thành đề tài ***"Multi-label movie genres classification from original overview"***, bên cạnh sự nỗ lực và cố gắng của cả nhóm, chúng em đã nhận được rất nhiều sự quan tâm, hướng dẫn và giúp đỡ quý báu từ quý Thầy/Cô và bạn bè.

Chúng em xin bày tỏ lòng biết ơn sâu sắc đến **TS. Nguyễn Trọng Chỉnh**, giảng viên hướng dẫn, người đã tận tình định hướng, hỗ trợ chuyên môn và góp ý cho nhóm trong suốt quá trình nghiên cứu – từ việc tìm hiểu lý thuyết về xử lý ngôn ngữ tự nhiên, lựa chọn mô hình, đến cách xây dựng phương pháp tiếp cận và xử lý vấn đề kỹ thuật.

Dù đã nỗ lực hết mình để hoàn thiện đề tài, nhưng do thời gian và kiến thức còn hạn chế, nhóm khó tránh khỏi những thiếu sót nhất định. Chúng em rất mong nhận được sự góp ý quý báu từ quý Thầy/Cô để có thể cải thiện và phát triển hơn trong các nghiên cứu tiếp theo.

Xin chân thành cảm ơn!

Nhóm Phông BART

MỤC LỤC

DANH MỤC HÌNH ẢNH VÀ BẢNG

Chương 1. GIỚI THIỆU BÀI TOÁN

1.1. Đề tài

Multilabel Movie Genres Classification from Original Overview (Phân loại đa lớp các thể loại phim từ mô tả tổng quan) là đề tài hướng đến bài toán phân loại nhiều lớp từ văn bản là một miêu tả tổng quan ngắn gọn của một bộ phim, áp dụng xử lý ngôn ngữ tự nhiên để phân tích sự tương quan và độ chính xác giữa miêu tả được đưa ra với thể loại và nội dung phim.

1.2. Mục tiêu đề tài

Với đề tài này, nhóm thực hiện việc phân loại với các mô hình ngôn ngữ BERT (Bidirectional Encoder Representations from Transformer) khác nhau như DistilBERT, BERT-base-cased, BERT-based-uncased cùng với mô hình máy học Logistic Regression kết hợp TF-IDF. Từ đó so sánh ưu nhược điểm của từng mô hình khi áp dụng cho bài toán phân loại đa nhãn. Mục tiêu của đề tài này nhằm tìm kiếm được mô hình phù hợp cho bài toán phân loại đa nhãn trong lĩnh vực phim ảnh, đánh giá sự tương quan giữa mô tả của bộ phim và thể loại phim, giúp cải thiện chất lượng và tính chính xác của mô tả đối với khán giả xem phim.

Chương 2. DỮ LIỆU

2.1. Nguồn gốc dữ liệu

- Dữ liệu được trích xuất từ bộ dữ liệu gốc wykonos/movies trên Hugging Face. Bộ dữ liệu gồm các thông tin như ID, tên phim, thể loại, ngôn ngữ, ngày ra mắt,... được viết ở ngôn ngữ tiếng Anh và lưu trữ ở định dạng CSV (Comma-separated Values).
- Link dataset: wykonos/movies[1]
- Bộ dữ liệu gốc bao gồm 20 cột và 722796 mẫu dữ liệu, đa dạng thông tin về từ quốc gia, thể loại, đơn vị sản xuất, tagline, ngày ra mắt, doanh thu,...
- Một vài mẫu ví dụ được lấy từ dataset gốc:

	id	title	genres	original_language	overview
0	385687	Fast X	Action-Crime-Thriller	en	Over many missions and against impossible odds...
1	758323	The Pope's Exorcist	Horror-Mystery-Thriller	en	Father Gabriele Amorth Chief Exorcist of the V...
2	640146	Ant-Man and the Wasp: Quantumania	Action-Adventure-Science Fiction	en	Super-Hero partners Scott Lang and Hope van Dy...
3	677179	Creed III	Drama-Action	en	After dominating the boxing world Adonis Creed...
4	502356	The Super Mario Bros. Movie	Animation-Family-Adventure-Fantasy-Comedy	en	While working underground to fix a water main ...
5	631842	Knock at the Cabin	Horror-Mystery-Thriller	en	While vacationing at a remote cabin a young gi...
6	603692	John Wick: Chapter 4	Action-Thriller-Crime	en	With the price on his head ever increasing Joh...
7	840326	Sisu	Action-War	fi	Deep in the wilderness of Lapland Aatami Korpi...
8	646389	Plane	Action-Adventure-Thriller	en	After a heroic job of successfully landing his...
9	569094	Spider-Man: Across the Spider-Verse	Action-Adventure-Animation-Science Fiction	en	After reuniting with Gwen Stacy Brooklyn’s ful...
10	505642	Black Panther: Wakanda Forever	Action-Adventure-Science Fiction	en	Queen Ramonda Shuri M’Baku Okoye and the Dora ...

Hình 2.1.1. Một vài dòng ví dụ minh họa trong bộ dữ liệu gốc ở 5 cột đầu tiên

	popularity	production_companies	release_date	budget	revenue	runtime	status
0	6682.100	Universal Pictures-Original Film-One Race-Perf...	2023-05-17	340000000.0	6.867000e+08	142.0	Released
1	5953.227	Screen Gems-2.0 Entertainment-Jesus & Mary-Wor...	2023-04-05	18000000.0	6.567582e+07	103.0	Released
2	4425.387	Marvel Studios-Kevin Feige Productions	2023-02-15	200000000.0	4.757662e+08	125.0	Released
3	3994.342	Metro-Goldwyn-Mayer-Proximity Media-Balboa Pro...	2023-03-01	75000000.0	2.690000e+08	116.0	Released
4	3859.926	Universal Pictures-Illumination-Nintendo	2023-04-05	100000000.0	1.278767e+09	92.0	Released
5	3422.537	Blinding Edge Pictures-Universal Pictures-Film...	2023-02-01	20000000.0	5.200000e+07	100.0	Released
6	2808.342	Thunder Road-87Eleven-Summit Entertainment-Stu...	2023-03-22	90000000.0	4.317692e+08	170.0	Released
7	2634.212	Subzero Film Entertainment-Good Chaos-Stage 6 ...	2023-01-27	6200000.0	1.056863e+07	91.0	Released
8	2618.646	MadRiver Pictures-Di Bonaventura Pictures-G-BA...	2023-01-12	25000000.0	5.100000e+07	107.0	Released
9	2550.738	Columbia Pictures-Sony Pictures Animation-Lord...	2023-05-31	100000000.0	5.126096e+08	140.0	Released
10	2525.408	Marvel Studios	2022-11-09	250000000.0	8.585356e+08	162.0	Released

Hình 2.1.2. Một vài dòng ví dụ minh họa trong bộ dữ liệu gốc ở 7 cột tiếp theo

	vote_average	vote_count	credits	keywords
0	7.331	1856.0	Vin Diesel-Michelle Rodriguez-Tyrese Gibson-Lu...	sequel-revenge-racing-family-cars
1	7.433	545.0	Russell Crowe-Daniel Zovatto-Alex Essoe-Franco...	spain-rome italy-vatican-pope-pig-possession-c...
2	6.507	2811.0	Paul Rudd-Evangeline Lilly-Jonathan Majors-Kat...	hero-ant-sequel-superhero-based on comic-famil...
3	7.262	1129.0	Michael B. Jordan-Tessa Thompson-Jonathan Majo...	philadelphia pennsylvania-husband wife relatio...
4	7.764	4042.0	Chris Pratt-Charlie Day-Anya Taylor-Joy-Jack B...	video game-gorilla-plumber-magic mushroom-anth...
5	6.457	888.0	Dave Bautista-Jonathan Groff-Ben Aldridge-Kris...	based on novel or book-sacrifice-cabin-faith-e...
6	7.904	3039.0	Keanu Reeves-Donnie Yen-Bill Skarsgård-Ian McS...	new york city-martial arts-hitman-sequel-organ...
7	7.393	261.0	Jorma Tommila-Aksel Hennie-Jack Doolan-Mimosa ...	world war ii-nordic mythology-lapland-finnish ...
8	6.901	785.0	Gerard Butler-Mike Colter-Yoson An-Tony Goldwy...	pilot-airplane-philippines-held hostage-plane ...
9	8.640	1684.0	Shameik Moore-Hailee Steinfeld-Brian Tyree Hen...	sacrifice-villain-comic book-sequel-superhero-...
10	7.338	3922.0	Letitia Wright-Lupita Nyong'o-Danai Gurira-Win...	loss of loved one-hero-sequel-superhero-based ...

Hình 2.1.3. Một vài dòng ví dụ minh họa trong bộ dữ liệu gốc ở 4 cột tiếp theo

	poster_path	backdrop_path	recommendations
0	/fiVW06jE7z9YnO4trhaMEdclSiC.jpg	/4XM8DUTQb3lhLemJC51Jx4a2EuA.jpg	19603-445954-697843-603692-781009-502356-74735...
1	/9JBEPLTPSm0d1mbEcLxULjJq9Eh.jpg	/hiHGRbyTcbZoLsYYkO4QiCLYe34.jpg	713704-296271-502356-1076605-1084225-1008005-9...
2	/qnqGbB22YJ7dSs4o6M7exTpNxPz.jpg	/m8JTwHFwX7I7JY5fPe4SjqejWag.jpg	823999-676841-868759-734048-267805-965839-1033...
3	/cvsXj3I9Q2iyyIo95AecSd1tad7.jpg	/5i6SjyDbDWqyun8klUuCxrlFbyw.jpg	965839-267805-943822-842942-1035806-823999-107...
4	/qNBAXBIQlnOThrVvA6mA2B5ggV6.jpg	/2klQ1z1fcHGgQPevbEQdkCnzyuS.jpg	713704-385687-640146-60898-758323-1008005-4935...
5	/dm06L9pxDOL9jNSK4Cb6y139rrG.jpg	/zWDMQX0sPaW2u0N2pJaYA8bVVaJ.jpg	1058949-646389-772515-505642-143970-667216-104...
6	/vZloFAK7NmvMGKE7VkF5UHaz0I.jpg	/1inZm0xxXrpRfN0LxwE2TXzyLN6.jpg	1098239-802401-24791-502356-385687-525644-1076...
7	/tELs0h3PPicRbsuu5cQ8UFcBQno.jpg	/94TlUEhuwv8PhdIADEvSuwPljS5.jpg	552688-713704-882569-296271-502356-605886-8689...
8	/qi9r5xBgcc9KTxlOLjssEbDgO0J.jpg	/9Rq14Eyrf7Tu1xk0Pl7VcNbNh1n.jpg	505642-758769-864692-631842-1058949-925943-758...
9	/8Vt6mWEReuy4Of61Lnj5Xj704m8.jpg	/4HodYYKEIsGOdinkGi2Ucz6X9i0.jpg	496450-667538-385687-603692-298618-447277-9765...
10	/sv1xJUazXeYqALzczSZ3O6nkH75.jpg	/xDMIl84Qo5Tsu62c9DGWhmPI67A.jpg	436270-829280-76600-56969-312634-1037858-238-5...

Hình 2.1.4. Một vài dòng ví dụ minh họa trong bộ dữ liệu gốc ở 3 cột cuối cùng

- Tuy nhiên, do đây là bài toán phân loại đa lớp từ các mô tả tổng quan, cho nên từ bộ dữ liệu gốc, ta chỉ lấy ra 2 cột cần thiết đó là ‘genres’ và ‘overview’. Bộ dữ liệu mới của bài toán bao gồm cột ‘overview’ là đoạn văn mô tả bộ phim và cột ‘genres’ là các thể loại đúng của bộ phim ngăn cách nhau bằng dấu ‘-’. Bộ dữ liệu mới cũng đã được xử lý loại bỏ các mẫu dữ liệu có mô tả hoặc thể loại là None.

	overview	genres
0	Over many missions and against impossible odds...	Action-Crime-Thriller
1	Father Gabriele Amorth Chief Exorcist of the V...	Horror-Mystery-Thriller
2	Super-Hero partners Scott Lang and Hope van Dy...	Action-Adventure-Science Fiction
3	After dominating the boxing world Adonis Creed...	Drama-Action
4	While working underground to fix a water main ...	Animation-Family-Adventure-Fantasy-Comedy
5	While vacationing at a remote cabin a young gi...	Horror-Mystery-Thriller
6	With the price on his head ever increasing Joh...	Action-Thriller-Crime
7	Deep in the wilderness of Lapland Aatami Korpi...	Action-War
8	After a heroic job of successfully landing his...	Action-Adventure-Thriller
9	After reuniting with Gwen Stacy Brooklyn’s ful...	Action-Adventure-Animation-Science Fiction

Hình 2.1.5. Một vài dòng ví dụ minh hoạ cho dataset được sử dụng trong bài toán

2.2. Định nghĩa các nhãn dữ liệu

- Bộ dữ liệu bao gồm 19 nhãn dữ liệu về nhiều thể loại phim khác nhau. Một mẫu có thể có nhiều nhãn do tính chất của phim ảnh bao gồm nhiều thể loại khác nhau được xác định thông qua nhiều yếu tố như bối cảnh, nội dung,...[2]

Nhãn	Định nghĩa
Comedy	Thể loại phim mang tính chất hài hước, gây cười cho người xem.
Fantasy	Thể loại phim mang bối cảnh không có thật, bao gồm nhiều yếu tố siêu nhiên.
Family	Thể loại phim về tình cảm gia đình, giữa con với ba mẹ, cháu với ông bà,...
Mystery	Thể loại phim mang các yếu tố bí ẩn, chưa có lời giải đáp.
Documentary	Thể loại phim tài liệu, ghi chép lại về một hiện tượng, sự kiện, nhân vật trong đời sống với các hình ảnh thực tế.
Western	Thể loại phim được lấy bối cảnh từ đời sống ở miền Tây nước Mỹ.
Drama	Thể loại phim mang nhiều yếu tố tâm lý, đánh vào cảm xúc, quan hệ và tình huống giữa người với người.
Music	Thể loại phim mà nhân vật thể hiện cảm xúc, lời nói qua các bài hát, thường có ít thoại hơn các loại phim khác.
Thriller	Thể loại phim mang lại tâm trạng kỳ vọng, hồi hộp, mong chờ, lo lắng cho người xem, thường có nhịp độ nhanh.
War	Thể loại phim nói về hoặc mô tả lại bối cảnh, sự kiện chiến tranh.
Science Fiction	Thể loại phim mang nhiều yếu tố dựa trên khoa học vượt hiện đại nhưng không có thực, mang nhiều yếu tố tương lai.
TV Movie	Thể loại phim truyền hình dài tập, thường được chiếu thường hay cách ngày trên TV.
Romance	Thể loại phim lãng mạn, mang nhiều yếu tố về tình yêu, tình cảm.
Animation	Thể loại phim hoạt hình, với các nét vẽ gần gũi với trẻ em,
Crime	Thể loại phim tội phạm, với bối cảnh là các hoạt động tội ác, thường có sự đối đầu giữa cảnh sát và kẻ phạm tội.

Adventure	Thể loại phim phiêu lưu, có bối cảnh là các chuyến du hành mạo hiểm, tìm kiếm một mục tiêu.
History	Thể loại phim về lịch sử, tái hiện lại một hoặc nhiều sự kiện đã xảy ra trong lịch sử.
Horror	Thể loại phim kinh dị, mang nhiều yếu tố tâm linh, ma quỷ hoặc đánh vào tâm lý sợ hãi của người xem.
Action	Thể loại phim hành động, gồm nhiều cảnh quay đối kháng, đánh nhau, đối đầu giữa thiện và ác.

Bảng 2.2.1. Bảng định nghĩa các nhãn dữ liệu

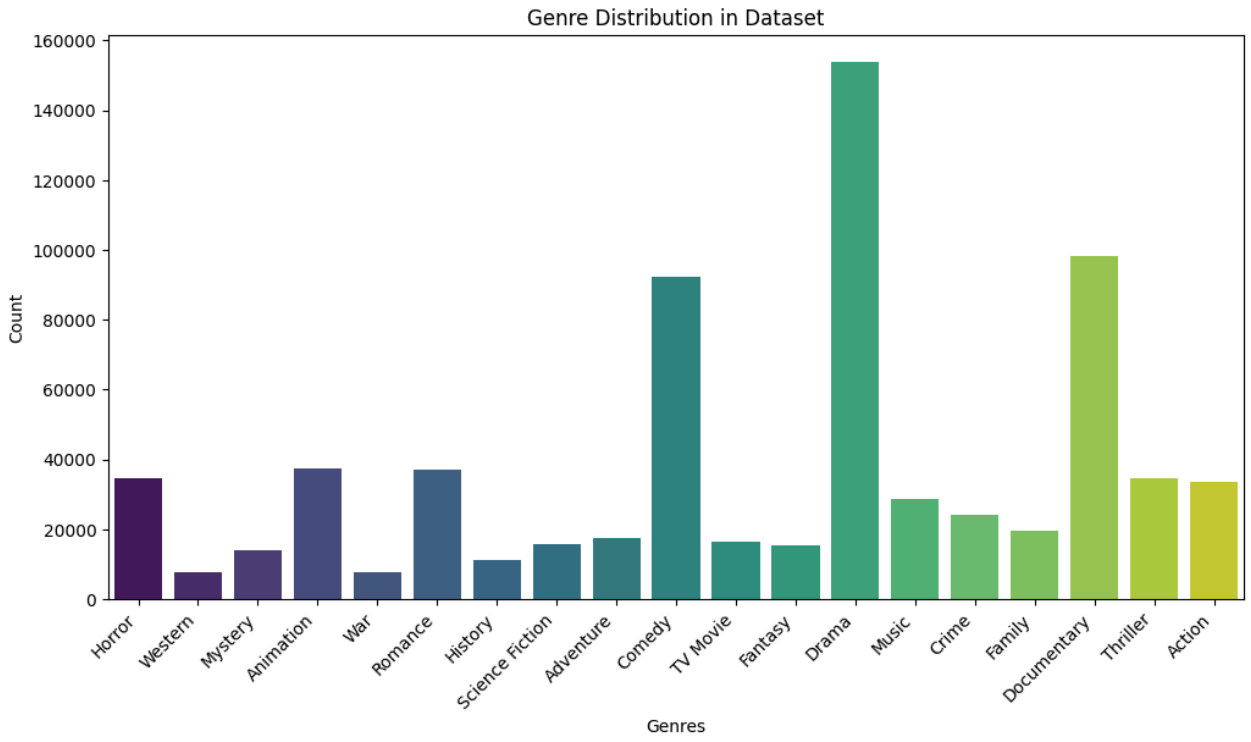
2.3. Thống kê dữ liệu

- Với tập dữ liệu trích xuất từ dataset gốc (wykonos/movies), nhóm có được các thông tin tiêu biểu về cột 'overview':

Max length	Min length	Average length	Number of vocabulary
180	5	47	293,923

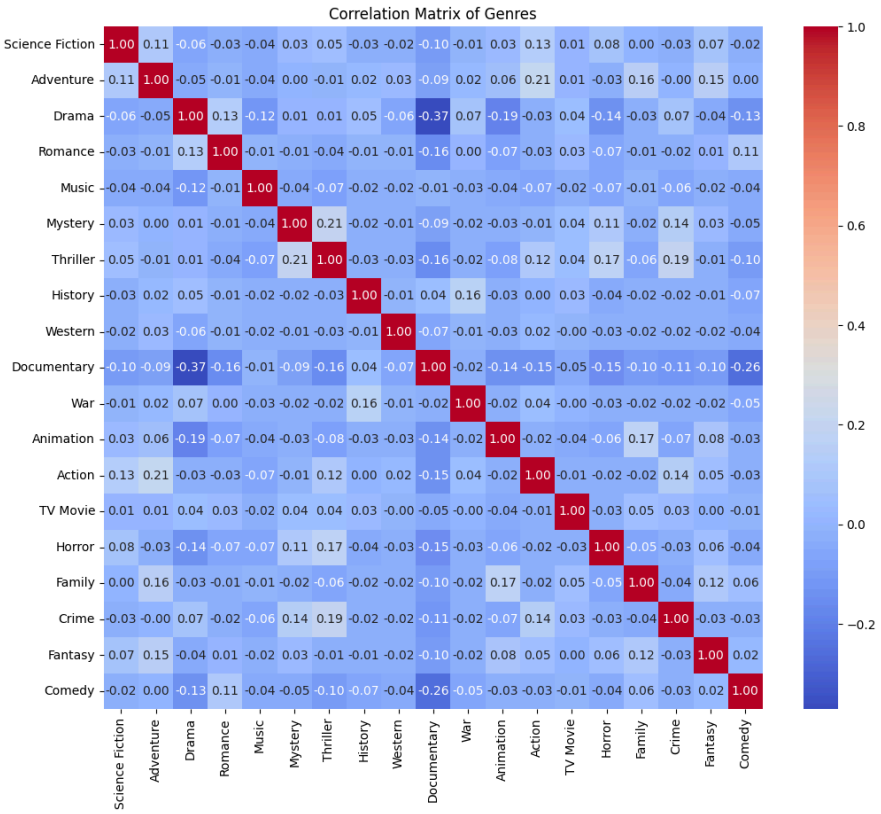
Bảng 2.3.1. Bảng thông tin về bộ dữ liệu dùng trong bài toán

- Phân bố các lớp:
  + Các mẫu dữ liệu chiếm phần lớn ở lớp Drama, theo sau đó là Documentary và Comedy với số lượng thấp hơn ⅔ so với lớp Drama.
  + Lớp có số lượng mẫu dữ liệu ít nhất là War với 7694 mẫu, theo sau đó là Western và History với 7701 và 11268 lần lượt.
  + Các lớp còn lại có số lượng gần tương đương nhau.



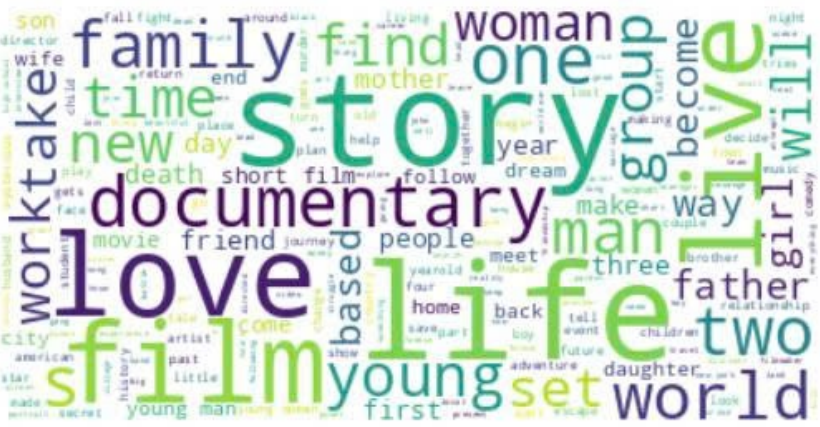
Hình 2.3.1 Phân bố của các lớp

- Ma trận tương quan giữa các lớp:
  + Các lớp ít có sự tương quan với nhau được thể hiện bằng việc chỉ số tương quan cao nhất là 0.21 với các cặp Mystery - Thriller và Action - Adventure.
  + Trong khi đó, lớp Documentary cho thấy ít tương quan với các lớp còn lại nhất, tiêu biểu ở Documentary - Drama với hệ số tương quan là -0.37.



Hình 2.3.2 Ma trận tương quan giữa các lớp

- Các từ vựng thường xuyên xuất hiện: Các từ "story", "love", "life" hay "film" có tần số xuất hiện cao, hợp lý do đây là các từ được sử dụng nhiều trong việc mô tả phim.



Hình 2.3.3 Tần suất xuất hiện của từ vựng

2.4. Ví dụ dữ liệu mẫu

Dữ liệu	Nhãn
Inspired by a true story an oddball group of cops criminals tourists and teens converge in a Georgia forest where a 500-pound black bear goes on a murderous rampage after unintentionally ingesting cocaine.	Thriller Comedy
Film tells about the person's way to freedom. About liberation from violence war and destruction of the surrounding world.	Animation
Jimmy Park is visiting home for the first time in years and has nothing to show for his time living overseas in Seoul. But old tensions come to a head when he confronts his homophobic sister at a family dinner.	Drama

Bảng 2.4.1. Ví dụ dữ liệu mẫu

2.5. Đánh giá dữ liệu

- Nhìn chung, dữ liệu và nhãn không hoàn toàn khớp với nhau, do một đoạn mô tả ngắn có thể không thể hiện hết được hoàn toàn thể loại của phim. Một số nhãn có thể khó trong việc nhận dạng, tiêu biểu là Animation, không thể hoàn toàn dựa vào dữ liệu để dự đoán có thuộc Animation hay không.

	Nội dung dữ liệu	Nhãn	Đánh giá
1	After a heroic job of successfully **landing** his **storm-damaged aircraft** in a **war zone** a **fearless** pilot finds himself between the agendas of multiple militias planning to take the plane and its passengers hostage.	Action Adventure Thriller	Đầu vào có nhắc đến landing và in a war zone. Cũng có các tính từ như fearless, Dễ dàng dự đoán là Action và Adventure. Tuy nhiên, vẫn chưa xác định được yếu tố của Thriller từ đoạn văn bản.
2	**A female sniper** on military leave promises to fulfill her fiancé's dying wish until she encounters a **hostile alien invasion** and is tasked with **saving countless lives.**	Action Science Fiction Fantasy	Các cụm "A female sniper", "Hostile alien invasion" hay "Saving countless lives", Dễ dàng đoán được là Science Fiction và Action. Tuy nhiên, vẫn chưa có yếu tố rõ ràng nào để xác định nhãn Fantasy.
3	An extraordinary young girl **discovers** her	Family	Từ các từ như "superpower" và

	superpower and **summons** the remarkable **courage** against all odds to help others change their stories whilst also taking charge of her own destiny. Standing up for what's right she's met with miraculous results.	Comedy Fantasy	"summons" có thể xác định đây là Fantasy. Các nhãn Comedy hay Family khó xác định được từ ngữ liệu. Có thể nhầm lẫn qua nhãn Adventure do có nhắc đến destiny, help others
4	Lucía a club dancer **on the run** takes **refuge** in a **sinister building** on the outskirts of Madrid where her sister Rocío lives with her daughter Alba.	Drama Horror Thriller	Các cụm "on the run" hay "sinister building" cho thấy có thể phim thuộc về Drama, Horror. Từ đầu vào, cũng có thể đoán được là Thriller, tuy nhiên vẫn chưa có từ khóa rõ ràng để xác định
5	When a **plane crashes at sea** dolphins rescue a little boy and raise him as family. He lives a carefree life **beneath the waves** until an **evil monster** seizes power over the **underwater world**. The boy is banished to dry land where a kind hearted captain takes him in. With the captain's help the boy sets out on a journey to **solve the mystery** of his true identity.	Animation Adventure Fantasy Comedy	Các nhãn Adventure và Fantasy có thể dễ dàng xác định nhờ vào "plane crashes at sea" hay "evil monster". Animation đoán được dựa vào "beneath the waves hay underwater world". Tuy nhiên chưa xác định được Comedy dựa vào ngữ liệu. Cũng như, ở đây, có nhắc đến "solve the mystery" nhưng lại không có nhãn mystery
6	Tanjiro Kamado joined with Inosuke Hashibira a boy **raised by boars** who **wears a boar's head** and Zenitsu Agatsuma a scared boy who reveals **his true power** when he sleeps boards the Infinity Train on a new mission with the Fire Hashira Kyojuro Rengoku to **defeat a demon** who has been **tormenting** the people and **killing** the **demon slayers** who oppose it!	Animation Action Adventure Fantasy Thriller	Có nhắc đến cụm "raised by boars", "wears a boar's head", Có thể là Animation và Fantasy. Nhãn Adventure, Action và Thriller cũng có thể xác định bằng "defeat a demon", "tormenting", "killing" hay "demon slayers"
7	After his retirement is interrupted by Gorr the God Butcher a **galactic killer** who seeks the extinction of the gods Thor Odinson enlists the help of King Valkyrie Korg and	Fantasy Action Comedy	Nhãn Action có thể được xác định qua "galactic killer" hoặc cũng có thể là "harrowing cosmic adventure", tuy nhiên, cụm này cũng có thể xác định

	ex-girlfriend Jane Foster who now wields Mjolnir as the Mighty Thor. Together they embark upon a **harrowing cosmic adventure** to **uncover** the **mystery** of the God Butcher's vengeance and stop him before it's too late.		thành nhãn Adventure, kèm theo từ "uncover". Cũng có nhắc đến "the mystery", nhưng lại không có nhãn mystery. Nhãn Fantasy có thể xác định qua Mighty Thor, là nhân vật thần thoại, tuy nhiên, cũng chưa rõ ràng
8	After more than thirty years of service as one of the Navy's top aviators and dodging the advancement in rank that would ground him Pete "Maverick" Mitchell finds himself training a detachment of TOP GUN graduates for **a specialized mission** the likes of which no living pilot has ever seen.	Action Drama	Dựa vào cụm "a specialized mission", có thể đoán được là Action và Drama. Dựa vào toàn bộ ngữ cảnh cũng khó có thể xác định được chắc chắn
9	**Undercover cop** Lucas White joins Vin Serento's LA gang of **illegal** street racers. They are fast and they are furious and they plan to double cross LA **crime** kingpin Juan Carlos de la Sol who hides his cash in a downtown Taco Bell. The gang's outrageous plan is as daring as it is **ridiculous** and will see them towing the whole restaurant at **crazy speeds.**	Comedy Action	Có nhắc đến "Undercover cop" hay "crime", dễ dàng xác định là Action. Các từ như ridiculous và cụm từ crazy speeds có thể xác định là Comedy. Tuy nhiên, "Undercover cop" và "crime" cũng rất khớp với thể loại Crime, nhưng nhãn Crime lại không được đánh
10	A new school year his **brother** Rodrick teases him over and over and **over and over again.** Will Greg manage to **get along with him?** Or will **a secret ruin everything?**	Animation Comedy Family	Dựa trên ngữ cảnh, không xác định được nhãn Animation. Có nhắc đến brother, có thể là Family. Nhãn Comedy có thể được xác định dựa vào "over and over and over again", tuy nhiên, cũng không chắc chắn. Nhãn Drama cũng có thể có ở đây, dựa vào câu dẫn ở cuối.
11	When a headstrong **street orphan** Seiya in search of his **abducted** sister unwittingly taps into hidden powers he **discovers** he might be the only person alive who can	Fantasy Action Adventure	Có thể xác định là Adventure dựa vào "street orphan", "abducted", "discover" và "embrace his destiny". Action có thể dựa vào "Knight of the

	protect a reincarnated **goddess** sent to watch over humanity. Can he let his past go and **embrace his destiny** to become a **Knight** of the Zodiac?		Zodiac". Fantasy có thể dựa trên từ "godless".
12	After a **young couple** moves into **a remote farmhouse** with their **infant son** the woman's struggles with **postpartum psychosis** begin to intensify... as the house reveals **secrets** of its own.	Horror Mystery Thriller	Do có cụm "postpartum psychosis", có thể dự đoán được là Horror, dựa vào "intensify" có thể đoán là Thriller. Từ secrets cũng có thể suy ra là Mystery, tuy nhiên, ngữ liệu cũng nhắc đến a young couple, remote farmhouse, infant son, nhưng lại không có nhãn Romance
13	A story of friendship between a young convict who is forced to work in a retirement home and a **group of crazy old people.** Together they **organize their escape**.	Comedy	Dựa vào cụm "a group of crazy old people", có thể dự đoán được nhãn Comedy, toàn bộ ngữ cảnh cho thấy cũng có yếu tố Drama, tuy nhiên không có nhãn Drama
14	When best friends and total opposites Debbie and Peter **swap homes** for a week they **get a peek into each other's lives** that could **open the door to love.**	Romance Comedy	Dựa vào hai cụm "get a peek into each other's lives" và "open the door to love", có thể đoán được thể loại là Romance, có nhắc đến "swap home", có thể dự đoán được là Comedy.
15	**A modern love story** set in **the near future** where an **AI building** is powered by human feelings. Due to a **software glitch** it **falls in love** with a real girl escapes the building into the body of a real man and tries to win her affections.	Comedy Romance Science Fiction	Có thể dự đoán được nhãn Comedy và Science Fiction dựa vào "A modern love story", "falls in love", "AI Building", "a software glitch". Tuy nhiên, chưa xác định được Comedy dựa vào ngữ cảnh
16	After 12 years **in prison** former high school football star Eddie Palmer **returns home** to put his life back together—and forms an unlikely bond with Sam an outcast boy from a troubled home. But Eddie's past **threatens to ruin** his new **life and family**.	Drama	Dựa vào các cụm "in prison", "return home" threatens to ruin" có thể dự đoán là Drama. Tuy nhiên, cũng có cụm "life and family", nhưng lại không có nhãn Family

17	After witnessing a **bizarre traumatic incident** involving a patient Dr. Rose Cotter starts **experiencing frightening occurrence**s that she can't explain. As an overwhelming terror begins taking over her life Rose must **confront her troubling past** in order to survive and escape her horrifying new reality.	Horror Mystery Thriller	Các cụm "bizarre traumatic incident" và "experiencing frightening occurrences" có thể thuộc thể loại Horror, Thriller và Mystery. Tuy nhiên, cũng có yếu tố Drama ở đây, dựa vào cụm "confront her troubling past", nhưng nhãn lại không có
18	Buster and his new cast now have their sights set on **debuting a new show** at **the Crystal Tower Theater** in glamorous Redshore City. But with no connections **he and his singers** must **sneak into the Crystal Entertainment offices** run by **the ruthless wolf mogul Jimmy Crystal** where **the gang** pitches the ridiculous idea of **casting the lion rock** legend Clay Calloway in **their show.** Buster must embark on a quest to find the now-isolated Clay and **persuade him to return to the stage.**	Animation Adventure Comedy Family	Dựa vào "the ruthless wolf mogul Jimmy Crystal", có thể phần nào xác định phim thuộc thể loại Animation. Xét tổng thể ngữ liệu, cũng cho thấy có yếu tố Adventure và Family. Cụm "sneak into the Crystal Entertainment offices" cho thấy cũng có Comedy, do từ ngữ sử dụng. Tuy nhiên, các cụm từ liên quan đến Music được nhắc đến rất nhiều, nhưng thể loại không thuộc Music
19	A short video- and sound montage by Lutz Mommartz.	None	Ngữ liệu có thể là miêu tả của một đoạn video ngắn, không phải phim, Không có thể loại
20	Tad accidentally **unleashes** an ancient spell **endangering** the lives of his friends Mummy Jeff and Belzoni. With **everyone against him** and only helped by Sara he sets off on an **adventure** to end the Curse of the Mummy.	Animation Adventure Comedy Family Fantasy	Từ "unleashes" và "adventure" cho thấy phim thuộc thể loại Adventure. Cụm "everyone against him" cũng cho thấy thuộc Family. Cách sử dụng từ ngữ, tên nhân vật như "Mummy Jeff" cho thấy phim có thể thuộc thể loại Fantasy và Comedy. Tuy nhiên, không thể xác định được phim thuộc Animation dựa trên ngữ liệu được cung cấp.
21	A group of soldiers are taken to the mountains of Wales to **investigate** a	Horror Fantasy	Cụm "strange looking monster" cho thấy phim thuộc thể loại Mystery và

	strange looking monster.	Mystery	Horror, tuy nhiên, vẫn có thể nhầm lẫn qua nhãn Action
22	A genetics specialist is sent to **stay in a coastal inn** to find the origin of an **infection** that has started spreading slowly.	Horror	Từ "infection" có thể cho thấy phim thuộc thể loại Science Fiction, tuy nhiên, không có nhãn này. Các nhãn cũng có khả năng như: Mystery, Adventure. Ngữ liệu chưa đủ để kết luận
23	The film is set in a small town where two modest couples find themselves caught up in a **game of life & death** with a **vicious gang** after **the discovery of a mysterious box** buried long ago.	Crime Thriller	Cụm "The discovery of mysterious box" cho thấy có thể thuộc thể loại Mystery, tuy nhiên, không được đánh nhãn. Ngữ liệu cũng không nhắc đến các yếu tố liên quan đến Crime. Các cụm "game of life & death" và "vicious gang"có thể thuộc thể loại Thriller
24	Tanjiro finds his **family** slaughtered and the lone survivor his sister **Nezuko Kamado** turned into a Demon. To his surprise however Nezuko still shows signs of human emotion and thought. Thus begins Tanjiro's journey to seek out the Demon who **killed their family** and turn his sister human again. A recap film of Kimetsu no Yaiba covering episodes 1-5 with extra footage.	Action Animation Fantasy	Cách đặt tên nhân vật là Demon có thể thuộc thể loại Action hoặc Horror. Nhân vật Nezuko Kamado giúp xác định thể loại Animation. Tổng quan cho thấy cũng có yếu tố Fantasy. Tuy nhiên, các từ liên quan đến family được nhắc đến nhiều, có thể gây nhầm lẫn đến nhãn Family
25	A reclusive English teacher suffering from severe **obesity** attempts to **reconnect** with his **estranged teenage daughte**r for one last chance at redemption.	Drama	Các từ/cụm như "obesity", "reconnect" hay "estranged teenage daughter" có thể giúp kết luận phim thuộc thể loại Drama.
26	Reclusive author Loretta Sage writes about exotic places in her popular **adventure novels** that feature a handsome cover model named Alan. While on tour promoting her new book with Alan Loretta gets	Action Adventure Comedy	Có các từ "adventure novels", "ancient city's lost treasure", phù hợp với nhãn Adventure. Các từ "kidnapped", "hero in real life", "rescue", phù hợp với nhãn

	kidnapped by an eccentric billionaire who hopes she can lead him to the **ancient city's lost treasure** that featured in her latest story. Alan determined to prove he can be a **hero in real life** and not just on the pages of her books sets off to **rescue** her.		Action. Tuy nhiên mô tả chưa có từ ngữ rõ ràng để gán với nhãn Comedy.
27	Your favorite cozy camping **anime** returns with a movie as the former members of the Outdoors Club get together again this time to build a campsite! Reunite with Nadeshiko Rin Chiaki Aoi and Ena as they gather around the campfire once more with good food and good company.	Animation Comedy	Có từ ngữ "anime", phù hợp với nhãn Animation. Tuy nhiên chưa có từ ngữ mang nghĩa rõ ràng phù hợp với nhãn Comedy.
28	Álex and Cata two young women who live on the outskirts of the city a desolate place full of empty stores and ramshackle bars where **dreams end up fading** have been **preparing their escape all their lives**.	Drama	Có cụm "dreams end up fading" và "preparing their escape all their lives", có khả năng phù hợp với nhãn Drama, nhưng vẫn có thể nhầm lẫn.
29	While under heavily armed guard the dangerous convicts aboard a cargo ship unite in a coordinated escape attempt that soon escalates into a **bloody all-out riot**. But as the fugitives continue their brutal campaign of terror they soon discover that not even the most vicious among them is safe from the **horror** they unknowingly unleashed from the **darkness** below deck.	Action Thriller Horror	Có cụm từ "bloody all-out riot", "horror", "darkness", phù hợp với nhãn Action và Horror. Câu thứ 2 trong mô tả mang cảm giác bí ẩn và kích thích sự tò mò, phù hợp với nhãn Thriller.
30	Through a series of unfortunate events three **mummies** end up in present-day London and embark on a **wacky and hilarious journey in search** of an old ring belonging to the Royal Family that was stolen by the ambitious archaeologist Lord Carnaby.	Adventure Animation Comedy Family Fantasy	Có cụm "journey in search", phù hợp với Adventure. Có từ "mummies", phù hợp với Fantasy. Có cụm từ "wacky and hilarious", phù hợp với Comedy. Tuy nhiên, có cụm "Royal Family" nhưng vẫn chưa đủ để xác định nhãn Family có phù hợp hay không.

31	Special agent Orson Fortune and his team of operatives recruit one of Hollywood's biggest movie stars to help them on an **undercover mission** when the sale of a **deadly new weapons** technology **threatens to disrupt the world order**.	Action Thriller Comedy	Có cụm "undercover mission" và "deadly new weapons", phù hợp với nhãn Action. Có cụm "threaten to disrupt the world order", Có thể phù hợp với nhãn Thriller. Không có từ ngữ phù hợp với nhãn Comedy.
32	The fearless one-eyed **weasel** Buck teams up with mischievous **possum** brothers Crash & Eddie as they head off on a new **adventure** into Buck's home: The Dinosaur World.	Animation Comedy Adventure Family	"weasel" và "possum", Do là động vật nên phải là phim hoạt hình, phù hợp với nhãn Animation. Từ "adventure" thể hiện rõ dấu hiệu cho nhãn Adventure". Chưa có từ ngữ thể hiện rõ ràng cho nhãn Comedy và Family.
33	To put their **demons** to rest once and for all Josh Lambert and a college-aged Dalton Lambert must go deeper into The Further than ever before facing their family's **dark past** and a host of new and more **horrifying terrors** that lurk behind the red door.	Horror	Các từ ngữ "demons", "dark past", "horrifying terrors" thể hiện rõ nhãn Horror.
34	In the follow-up to "The Battle At Lake Changjin" brothers Wu Qianli and Wu Wanli undertake a new task for the **People's Volunteer Army** defending a bridge part of the American **troops**' escape route from the advancing Chinese.	War History Action Drama	Nội dung có nhắc đến quân đội thông qua các từ "Army", "troops", phù hợp với nhãn War và Action. Có nhắc đến "People's Volunteer Army" là một trận chiến trong lịch sử nên phù hợp với nhãn History. Không có từ nào quá phù hợp để miêu tả nhãn Drama.
35	Despite his **family**'s baffling generations-old ban on **music** Miguel dreams of becoming an accomplished **musician** like his idol Ernesto de la Cruz. Desperate to prove his talent Miguel finds himself in the stunning and colorful **Land of the Dead** following a mysterious chain of events. Along the way he meets charming trickster Hector and together they	Family Animation Fantasy Music Comedy Adventure	Có "family", phù hợp với nhãn Family. Cụm từ "Land of the Dead" là một địa danh không có thật, thể hiện rõ cho nhãn Fantasy. Các từ "music", "musician", phù hợp với nhãn Music. Nội dung về hành trình của Miguel ở Land of the Dead phù hợp với nhãn Adventure. Chưa có từ ngữ phù hợp với nhãn Animation và

	set off on an extraordinary journey to unlock the real story behind Miguel's family history.		Comedy.
36	Twenty-five years after a streak of **brutal murders** shocked the quiet town of Woodsboro a new killer has donned the Ghostface mask and begins targeting a group of teenagers to resurrect secrets from the town's deadly past.	Horror Mystery Thriller	Cụm "brutal murders", "killer" thể hiện rõ nhãn Horror. Nội dung mang lại cảm giác tò mò, bí ẩn về câu chuyện chưa được giải quyết, phù hợp với nhãn Mystery và Thriller.
37	The president of a farmers' association wants to set up a community farming initiative and takes on a big shot who wants to destroy his plans so that he can start a bio-diesel project on the land.	Drama Action	Mô tả không có từ ngữ hay nội dung quá phù hợp với cả nhãn Drama và Action.
38	A reunion between two estranged sisters gets cut short by the rise of flesh-possessing demons thrusting them into a primal battle for survival as they face the most nightmarish version of family imaginable.	Animation Science Fiction Adventure	Mô tả không có từ ngữ hay nội dung quá phù hợp với cả nhãn Animation và Adventure. Nội dung có nhắc về cuộc chiến với ác quỷ, có thể phù hợp với nhãn Science Fiction.
39	Truman Burbank is the star of The Truman Show a 24-hour-a-day reality TV show that broadcasts every aspect of his life without his knowledge. His entire life has been an unending **soap opera** for consumption by the rest of the world. And everyone he knows including his wife and his best friend is really an actor paid to be part of his life.	Comedy Drama	Cụm từ "soap opera" thể hiện rõ tính chất của nhãn Drama. Tuy nhiên chưa có từ ngữ và nội dung phù hợp với nhãn Comedy.
40	The eleventh installment in The Fast Saga.	Action Crime	Mô tả không có từ ngữ hay nội dung phù hợp với cả hai nhãn Action và Crime.
41	Simba idolizes his **father** King Mufasa and takes to heart his own royal destiny. But not everyone in the kingdom celebrates the new cub's arrival. Scar Mufasa's **brother**—and	Adventure Drama Family	Có các từ ngữ "father", "brother" phù hợp với nhãn Family. Có cụm từ "betrayal tragedy and drama" thể hiện rõ nhãn Drama. Nội dung của câu

	former heir to the throne—has plans of his own. The battle for Pride Rock is ravaged with **betrayal tragedy and drama** ultimately resulting in Simba's exile. With help from a curious pair of newfound friends Simba will have to figure out how to grow up and take back what is rightfully his.		cuối về việc học cách lớn lên có thể là dấu hiệu của nhãn Adventure.
42	At last it's Wembley! The milestone **concert** of 2 hours and 30 minutes that filled the immense stadium returns to ARMY all around the world!	Music	Chứa từ ngữ "concert" phù hợp với nhãn Music.
43	Unhappily **married** aristocrat Lady Chatterley begins a torrid affair — and **falls deeply in love** — with the gamekeeper on her husband's country estate.	Romance Drama	Có chứa từ ngữ "married", "affair", "falls deeply in love" phù hợp với nhãn Romance. Nội dung về tình tay ba phù hợp với nhãn Drama.
44	In a time when **monsters** walk the Earth humanity's fight for its **future** sets Godzilla and Kong on a collision course that will see the **two most powerful forces of nature on the planet collide** in a spectacular battle for the ages.	Action Fantasy Science Fiction	Cụm "two most powerful forces … collide" phù hợp với nhãn Action. Có nhân vật là "monster" nên phù hợp với nhãn Fantasy. Có từ "future" and nội dung không thực nên hợp với nhãn Science Fiction.
45	Creatively unfulfilled and facing financial ruin Nick Cage must accept a $1 million offer to attend the birthday of a dangerous superfan. Things take a wildly unexpected turn when Cage is **recruited by a CIA operative** and forced to live up to his own legend channeling his most iconic and beloved on screen characters in order to save himself and his loved ones.	Action Comedy Crime	Cụm từ "recruited by a CIA operative" có thể phù hợp với nhãn Action. Nội dung của mô tả nhắc về số tiền lớn và các điều kiện phù hợp với nhãn Crime. Chưa có từ ngữ hay nội dung phù hợp với nhãn Comedy.
46	On the run with a bag full of cash after a robbing his former **crime boss**—and a potentially fatal wound—Freddy slips onto	Crime Thriller	Nội dung nói về việc cướp tiền và từ ngữ "crime boss" phù hợp với nhãn Crime. Có cụm "life slipping through

	a bus headed into the unrelenting California desert. With his **life slipping through his fingers** Freddy is left with very **few choices to survive.**		his fingers", "few choices to survive" phù hợp với nhãn Thriller.
47	Armed with the astonishing ability to shrink in scale but increase in strength master thief Scott Lang must embrace his inner-hero and help his mentor Doctor Hank Pym protect the secret behind his spectacular Ant-Man suit from a new generation of towering threats. Against seemingly insurmountable obstacles Pym and Lang must plan and pull off a heist that will **save the world**.	Science Fiction Action Adventure	Nội dung có nhắc về các khả năng siêu phàm phù hợp với nhãn Science Fiction. Có cụm "save the world" phù hợp với nhãn Action. Chưa có từ ngữ nội dung quá phù hợp với nhãn Adventure.
48	Caleb a former government assassin in hiding who resurfaces when his protégé the equally deadly killer known as Banshee discovers a bounty has been placed on Caleb's head.	Thriller Action	Nội dung nói về sự đối đầu giữa 2 sát thủ phù hợp với cả 2 nhãn Thriller và Action.
49	When the Bad Guys a **crew of criminal animals** are finally caught after years of **heists** and being the world's most-wanted **villains** Mr. Wolf brokers a deal to **save them all from prison**.	Animation Action Adventure Crime Comedy	Có từ "crew of criminal animal" nên phù hợp với nhãn Animation. Có từ ngữ "criminal", "heists", "villains" phù hợp với nhãn Crime. Nội dung có "save from prison", có thể phù hợp với nhãn Adventure. Chưa có từ ngữ và nội dung phù hợp với Comedy.
50	A recently **married** scholar **goes on a quest** for knowledge of other people's wives based on his philosophical differences with the Sack Monk. He encounters the Flying Thief who agrees to help him find women but only if he attains a penis as big as a horse's. The scholar has a surgeon attach said unit and he's off and running on his mission only to find that there are obstacles to his new lifestyle such as jealous	Drama Adventure Comedy Romance	Nội dung mang nhiều tình tiết tréo ngoe phù hợp với nhãn Drama. Cụm "goes on a quest" phù hợp với nhãn Adventure. Nội dung lạ kỳ mang một chút hài hước phù hợp với nhãn Comedy. Có các cụm từ "married", "husband" phù hợp với nhãn Romance.

	husbands and treacherous females.		
51	An LA **vampire hunter** has a week to come up with the cash to pay for his kid's tuition and braces. Trying to make a living these days just might kill him.	Action Fantasy Horror Comedy	Có cụm "vampire hunter" cho thấy phim thuộc thể loại Horror và Action. Yếu tố Fantasy cũng thể hiện qua "vampire" là sinh vật không có thật, mang tính siêu nhiên. Không chứa từ ngữ mang yếu tố Comedy.
52	A mother daughter and two of their friends **live in an old Western-style house** in **Suginami**. As **secrets** are revealed they face **challenges together**.	Romance TV Movie	Yếu tố về địa điểm như "Sugunami" và "live in an old Western-style house" gợi lên một khung cảnh đời sống thường ngày, ngoài ra còn có yếu tố tình tiết nhẹ nhàng, tập trung vào cuộc sống, bí mật và thử thách cùng nhau → Phù hợp với nhãn TV Movie. Không có từ ngữ thể hiện cho nhãn Romance. Dễ nhầm lẫn sang các yếu tố khác như: Family, Drama
53	Determined to **prove herself Officer Judy Hopps** the first **bunny** on Zootopia's police force jumps at the chance to **crack her first case** - even if it means partnering with **scam-artist fox Nick Wilde** to **solve the mystery.**	Animation Adventure Family Comedy	Có yếu tố động vật được nhân hóa thành con người như "bunny", "fox" → đặc trưng của nhãn Animation. Các cụm từ như "prove herself", "crack her first case", "solve the mystery" thể hiện tuyến câu chuyện phiêu lưu phá án, giải đáp bí ẩn → đặc trưng của nhãn Adventure. Yếu tố Comedy có thể được thể hiện qua mối quan hệ giữa "Officer Judy Hopps" và "scam-artist fox Nick Wilde". Chưa có từ ngữ thể hiện cho nhãn Family.
54	Set in a **post-apocalyptic** world young	Action	Có cụm từ thế giới hậu tận thế

	Thomas is deposited in a community of boys after **his memory is erased soon** learning they're all **trapped in a maze** that will require him to join forces with fellow **"runners"** for a shot at **escape**.	Mystery Science Fiction Thriller	"post-apocalyptic" → Phù hợp với nhãn Science Fiction. Có yếu tố Action và Thriller thông qua các từ ngữ như "trapped", "escape", "runners". Motif phổ biến trong Science Fiction và Mystery: mất trí nhớ "his memory is erased soon", bị nhốt trong mê cung "trapped in a maze".
55	How many young people are planning to set up a restaurant? One of them **has died** and a **significant amount of money is missing**.	Drama	Có câu hỏi tu từ tăng tính nghiêm túc, Các cụm từ như "has died", "significant amount of money is missing" gợi lên yếu tố nghi ngờ và xung đột giữa các cá nhân → Đặc trưng của thể loại Drama.
56	A group of **Bulgarian soldiers** go on a **mission** during **the Balkan War**.	Drama War	Các cụm từ "Bulgarian soldiers", "mission", "the Balkan War" thể hiện rõ về đặc trưng của nhãn War. Chưa có yếu tố nào để xác định đây là nhãn Drama.
57	Spanning the years 1945 to 1955 a chronicle of the fictional Italian-American Corleone **crime family**. When organized crime family patriarch Vito Corleone barely survives an attempt on his life his youngest son Michael steps in to take care of the would-be **killers** launching a campaign of **bloody revenge**.	Drama Crime	Các yếu tố như "crime family", "killers", "bloody revenge" thể hiện rõ nhãn Crime. Câu chuyện dài, có sự xung đột "bloody revenge", "an attempt on his life" thể hiện nhãn Drama.
58	Batman raises the stakes in his **war on crime**. With the help of Lt. Jim Gordon and District Attorney Harvey Dent Batman sets out to dismantle the remaining **criminal organizations** that plague the streets. The partnership proves to be effective but they soon find themselves **prey to** a **reign of**	Drama Action Crime Thriller	"war on crime", "criminal organizations", "dismantle", "rising criminal mastermind" — đều là từ khóa mạnh của Crime và Action. "reign of chaos", "terrified citizens", và "prey to" → gợi cảm giác căng thẳng, nguy hiểm, không chắc chắn

	chaos unleashed by a **rising criminal mastermind** known to the **terrified citizens** of Gotham as the Joker.		— những yếu tố đặc trưng của Thriller. Drama được gán nhưng không có cụm từ nào gợi tả chi tiết.
59	The aftermath of a shocking **explosion at the Chernobyl nuclear power station** made hundreds of people **sacrifice their lives** to clean up the site of the catastrophe and to successfully **prevent an even bigger disaster** that could have turned a large part of the European continent into an uninhabitable exclusion zone. This is their story.	Drama History Adventure Action Fantasy Mystery	Những cụm từ "sacrifice their lives", "prevent an even bigger disaster", "this is their story" gợi lên yếu tố bi kịch con người, xung đột giữa trách nhiệm và cái chết, cảm xúc hi sinh – đặc trưng rõ ràng của Drama. “explosion at the Chernobyl nuclear power station” là một sử kiện có thật trong lịch sử → nhãn History. Các nhãn còn lại như Adventure, Action, Fantasty và Mystery chưa được ngữ liệu thể hiện rõ.
60	**Maleficent** and **her goddaughter Aurora** begin to question the **complex family ties** that bind them as they are pulled in different directions by impending **nuptials** **unexpected allies** and **dark new forces** at play.	Family Fantasy Adventure Action	Các nhãn Fantasy và Family có thể dễ dàng xác định nhờ vào “Maleficent”, “dark new forces”, và “complex family ties”. Nhãn Adventure có thể được suy luận từ “unexpected allies” và xung đột định hướng, tuy nhiên không có từ khóa cụ thể về hành trình hoặc trận chiến. Nhãn Action chưa thể xác định được từ ngữ liệu trong overview này.
61	When a lively young family moves in next door **grumpy widower** Otto Anderson meets his match in a **quick-witted** pregnant woman named Marisol leading to an **unlikely friendship** that **turns his world upside down**.	Comedy Drama	Các nhãn Comedy và Drama có thể dễ dàng xác định từ cụm như “grumpy widower”, “quick-witted”, “unlike friendship” và “turns his world upside down”. Comedy được hỗ trợ bởi mô típ đối lập tính cách tạo tình huống hài, trong khi Drama thể hiện qua hành trình cảm xúc của nhân vật góa vợ tìm lại

			kết nối và thay đổi cuộc sống.
62	After the **bizarre death** of their father two siblings must **uncover the root cause** of their mother's **sudden paranormal terror** to save her.	Mystery Horror Drama	Các cụm từ "bizarre death" và "sudden paranormal terror" thể hiện rõ nhãn Horror. Tình tiết "uncover the root cause" tạo sự bí ẩn nên phù hợp với nhãn Mystery. Chưa có cụm từ nào thể hiện rõ nhãn Drama trong đoạn tóm tắt.
63	New recruit Justin Rosa must monitor **arms-smuggling cartel** member Eddie Flynn — and keep him alive at all costs. When a **SWAT team descends** on Flynn's home Rosa breaks protocol and contacts **the gangster** directly to save his life. As gunmen break into the Wire Room and chaos erupts Mueller and Rosa make a final desperate stand against the **corrupt agents** and officials who seek to **destroy evidence and kill** them both.	Action Crime Thriller	Nhãn Action, Thriller có thể dễ dàng xác định nhờ các cụm từ như "SWAT team descends", "arms-smuggling cartel", "corrupt agents", và "final desperate stand". Nhãn Crime có thể xác định nhờ các cụm từ "chaos", "destroy evidence and kill" và "the gangster". Ba nhãn này đều được hỗ trợ tốt từ từ khóa miêu tả chiến đấu, tội phạm và căng thẳng sinh tử.
64	In a post-apocalyptic future where **population control** is dictated by a high-school aptitude test two students discover the test is **smoke and mirrors** hiding a **larger conspiracy**.	Thriller Drama Horror	Những từ "population control", "smoke and mirrors" và "larger conspiracy" gợi lên sự căng thẳng, ẩn giấu thông tin, và phát hiện nguy cơ tiềm ẩn – là đặc trưng cốt lõi của thể loại Thriller. Nhãn Drama có thể được suy luận do bối cảnh học đường và xã hội, nhưng overview không nêu rõ chiều sâu cảm xúc nên chưa thể xác định chắc chắn. Nhãn Horror không có cơ sở vì không có dấu hiệu kinh dị trong ngữ liệu.
65	A mother and son find themselves faced with a **brutal alien invasion** where **survival** will depend on discovering the	Science Fiction Action	Nhãn Science Fiction có thể dễ dàng xác định từ cụm "alien invasion" và "truth about the enemy", cho thấy rõ

	unthinkable truth about the enemy.		yếu tố người ngoài hành tinh và giả tưởng khoa học. Nhãn Action cũng có thể suy ra từ cụm "brutal invasion" và "survival", hàm ý đến tình huống nguy hiểm và hành động phản kháng.
66	A **special bond** develops between plus-sized inflatable **robot Baymax** and prodigy **Hiro Hamada** who **team up** with a group of friends to form a band of **high-tech heroes**.	Adventure Family Animation Action Comedy	Các cụm từ "robot Baymax", " high-tech heroes" và "Hiro Hamada" có thể bổ sung cho nhãn Animation. Nhãn Family được thể hiện qua sự liên kết của các nhân vật "special bond", "team up". Các nhãn Adventure, Action và Comedy chưa được thể hiện rõ ràng qua overview này.
67	A beautiful strong-willed young royal refuses to wed the cruel sociopath to whom she is betrothed and is kidnapped and locked in a remote tower of her father's **castle**. With her scorned vindictive suitor intent on taking her father's **throne** the princess must **protect her family** and **save the kingdom**.	Fantasy Action	Nhãn Fantasy có thể dễ dàng xác định từ các cụm như "castle", "throne", và "kingdom", thể hiện rõ bối cảnh hoàng gia và phong cách cổ tích đặc trưng của Fantasy. Nhãn Action được hỗ trợ từ ngữ cảnh "protect her family" và "save the kingdom", ngụ ý hành động chống lại kẻ thù để bảo vệ ngai vàng.
68	Igor Grom is a **skilled policeman** from St. Petersburg known for his daring nature and **uncompromising attitude towards the criminals of all kinds**. Incredible **strength analytical mind and integrity** – these qualities make Major Grom the perfect policeman. Working tirelessly he always pushes through and **meets the challenges** standing in the way.	Action Adventure	Motif thường thấy trong thể loại Action: "skilled policeman", "uncompromising attitude towards the criminals of all kinds" Yếu tố Adventure thể hiện qua việc gặp nhiều thử thách "meets the challenges"
69	**Raquel's longtime crush on her next-door neighbor** turns into something more when	Drama Romance	Tình tiết "Raquel's longtime crush on her next-door neighbor" cho thấy

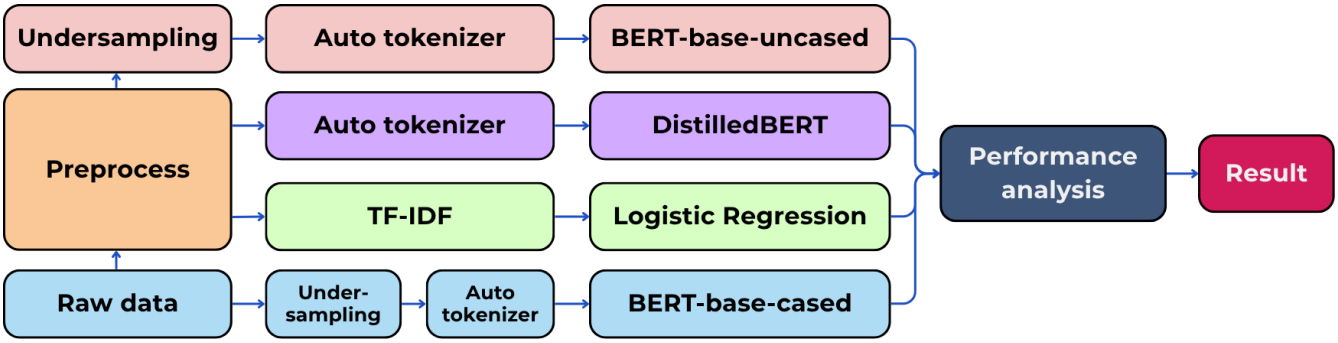
	he starts **developing feelings** for her despite his **family's objections**.		rằng có yếu tố tình cảm → Phù hợp với Romance. Sự ngăn cấm của gia đình "family's objections" và sự phát triển cảm xúc "developing feelings" cho thấy có yếu tố Drama.
70	**Ryder and the pups** are **called to Adventure City** to **stop Mayor Humdinger** from turning the bustling metropolis into a state of chaos.	Animation Family Adventure Comedy	Nhãn Animation có thể xác định nhờ vào tên nhân vật "Ryder and the pups" và động vật được nhân hóa. Nhãn Adventure được xác định từ cụm "called to Adventure City" và hành động "stop Mayor Humdinger". Nhãn Family và Comedy chưa được thể hiện rõ qua ngữ liệu.
71	Using **massive piloted robots** to **combat the alien threat** earth's survivors **take the fight** to the **invading alien** force lurking in the depths of the **Pacific Ocean**. Nearly defenseless in the face of the **relentless enemy** the forces of mankind have no choice but to turn to two unlikely heroes who now stand as earth's **final hope** against the mounting **apocalypse**.	Action Science Fiction Adventure	Chứa các từ khóa mạnh "combat the alien threat", "take the fight", "relentless enemy", "final hope" thể hiện rõ nhãn Action. Nhãn Science Fiction thể hiện qua "massive piloted robots", "alien threat", "Pacific Ocean", "the envading ailen" hay "apocalypse". Nhãn Adventure không được thể hiện quá rõ nét qua ngữ liệu.
72	Three friends decide to play The Red Book game. What they don't know is that in the house **evil** is waiting to be released. A long time ago a **witch** died as part of a **satanic ritual**. Every question they ask they get closer and closer to her. The book will reveal the answers to all your questions… but what if your host is an evil entity with a **thirst for blood?**	Horror	Có các cụm từ đặc trưng cho nhãn Horror như: "evil", "witch", "thirst for blood" Có thể gán thêm nhãn Mystery vì ngữ liệu có chứa các yếu tố siêu nhiên, không có thật như quỷ "evil" hay nghi lễ "satanic ritual".
73	After being bitten by a **genetically altered spider** at Oscorp nerdy but endearing high	Fantasy Action	Nhãn Action có thể dễ dàng xác định từ cụm "amazing powers" và

	school student Peter Parker is **endowed with amazing powers** to become the **superhero** known as Spider-Man.		"superhero", vốn là yếu tố hành động quen thuộc của dòng phim siêu anh hùng. Nhãn Fantasy cũng được hỗ trợ từ cụm "genetically altered spider" và "endowed with amazing powers", cho thấy sự xuất hiện của năng lực vượt tự nhiên.
74	**A dispute for the inheritance** of the father of this family **is the motor for this comedy** where siblings Andrea and Hector start a discussion with their aunt who turns out to be the universal heiress of their father's because they **don't comply with the rules** established by their father.	Comedy	Nhãn Comedy được xác định rõ ràng từ chính câu "is the motor for this comedy", cùng với các tình huống mâu thuẫn nhẹ mang tính gia đình như "dispute for the inheritance" và "don't comply with the rules".
75	**Carl Fredricksen** spent his entire life dreaming of **exploring the globe and experiencing life** to its fullest. But at age 78 life seems to have passed him by until a twist of fate (and a persistent **8-year old Wilderness Explorer named Russell**) gives him a **new lease on life**.	Animation Comedy Family Adventure	Nhãn Adventure thể hiện qua các cụm từ "exploring the globe and experiencing life" và "exploring the globe and experiencing life". Nhãn Animation thể hiện qua tên các nhân vật "Carl Fredricksen", "8-year old Wilderness Explorer named Russell". Hai nhãn Comedy và Family chưa được thể hiện rõ qua ngữ liệu này.

Bảng 2.5.1. Bảng phân tích dữ liệu

Chương 3. PHƯƠNG PHÁP NGHIÊN CỨU ĐỀ TÀI

3.1. Sơ đồ hóa phương pháp nghiên cứu đề tài



Hình 3.1.1. Sơ đồ quy trình thực hiện

Trước tiên, để thực hiện đề tài, nhóm đã chia thành các bước như trên sơ đồ:

- Đầu tiên, thực hiện các phương pháp tiền xử lý dữ liệu, do bộ dữ liệu có sẵn không hoàn toàn sạch. Như biểu đồ hình 2.3.1, phân bố giữa các lớp chưa thực sự đều, do đó có thể ảnh hưởng đến quá trình huấn luyện mô hình và dẫn đến việc học quá nhiều trên lớp này nhưng quá ít trong lớp khác, gây ra hiện tượng sai lệch trong quá trình dự đoán. Vì vậy, nhóm tiếp tục thử cắt giảm một phần dữ liệu để cân bằng phân phối giữa các lớp.
- Nhóm quyết định huấn luyện thử các tổ hợp khác nhau để đánh giá:
  + Về mô hình, nhóm chọn các mô hình:
    - BERT-base-uncased,
    - BERT-base-cased,
    - DistillBERT,
    - Multioutput Logistic Regression.
  + Về dữ liệu huấn luyện, nhóm chọn các dữ liệu:
    - Dữ liệu thô được cắt,
    - Dữ liệu đã tiền xử lý, nhưng không cắt,
    - Dữ liệu đã tiền xử lý và cắt.
  + Từ các dữ liệu và mô hình, nhóm có các tổ hợp:
    - Huấn luyện dữ liệu được tiền xử lý và cắt trên BERT-base-uncased,
    - Huấn luyện dữ liệu được tiền xử lý nhưng không cắt trên DistillBERT và Multioutput Logistic Regression,
    - Huấn luyện dữ liệu thô đã được cắt trên BERT-base-cased.

+ Đối với phương pháp vector hóa:
    - Sử dụng AutoTokenizer trong thư viện transformer đối với các kiến trúc thuộc họ BERT,
    - Sử dụng TF-IDF được hỗ trợ bởi scikit-learn đối với Logistic Regression, đồng thời kết hợp sử dụng MultiOutputClassifier của thư viện scikit-learn nhằm giải quyết bài toán đa đầu ra trên Logistic Regression.

- Từ các kết quả huấn luyện, nhóm tiến hành so sánh và đánh giá giữa các mô hình, từ đó chọn ra mô hình tốt nhất để giải quyết bài toán này.

3.2. Bài toán phân loại đa nhãn

Bài toán phân loại đa nhãn (multi-label classification) là một trong ba bài toán classification, tuy nhiên lại không được phổ biến như nhóm bài toán còn lại. Bài toán phân loại đa nhãn là bài toán mà một mẫu dữ liệu có thể bao gồm nhiều nhãn khác nhau, có thể áp dụng trong lĩnh vực Computer Vision với các bài toán nhận dạng hoặc lĩnh vực NLP với bài toán gán nhãn cho bài viết, gán nhãn thể loại cho bộ phim.

3.3. Các phương pháp đánh giá mô hình

3.3.1. Hàm loss BCEWithLogitsLoss

- Hàm loss BCEWithLogitsLoss là hàm được cung cấp bởi thư viện PyTorch. Hàm loss này kết hợp một lớp hàm Sigmoid và chỉ số Binary Cross Entropy Loss trong cùng một lớp. Hàm loss này ổn định hơn về mặt số học khi so với chỉ sử dụng một hàm Sigmoid đơn giản theo sau là Binary Cross Entropy Loss. Công thức tính toán của hàm này có thể biểu diễn như sau[3]:

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, \quad l_n = -w_n \left[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n)) \right],$$

- Hàm BCEWithLogitsLoss được sử dụng để đo lường độ lỗi giữa mẫu đã được biến đổi và mẫu gốc, ví dụ như sau khi sử dụng auto-encoder để biến đổi dữ liệu.
- Phương pháp đánh giá này thường được sử dụng cho bài toán phân loại nhị phân (binary classification). Tuy nhiên đối với bài toán multilabel, hàm loss này sẽ xử lý bài toán theo hướng one-vs-all, mỗi lớp sẽ được coi là một bài toán phân loại nhị phân riêng và tổng hợp bằng cách lấy tổng hoặc trung bình độ lỗi của mỗi lớp.

3.3.2. F1-Score

- F1 score là kết quả cho độ hòa hợp giữa precision và recall. F1 score cân bằng hai độ đo trên thành một con số đặc biệt quan trọng để theo dõi độ trade-off giữa chúng, cũng như cung cấp thêm một góc nhìn sâu hơn về độ hiệu quả của mô hình.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- Đối với bài toán phân loại đa lớp, F1 score được tính toán cho từng lớp riêng biệt. Vì vậy, kết quả cuối cùng cho cả mô hình được tính toán theo 2 hướng khác nhau:
  + Macro Average: Đối với hướng tính toán này, kết quả là trung bình F1 của các lớp mà không đánh trọng số cho từng kết quả. Độ đo này xem tất cả các lớp đều quan trọng như nhau.
  + Weighted Average: Với hướng tính toán này, kết quả là trung bình F1 của các lớp có đánh trọng số. Ở đây, trọng số được tính dựa trên số lượng mẫu đúng trong lớp, giải thích cho độ mất cân bằng giữa các lớp.

3.3.3. Jaccard score

- Jaccard score là đơn vị giúp đo độ tương đồng giữa 2 tập hợp thông qua giao và hợp của chúng. Jaccard score có công thức như sau[4]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Jaccard score tìm tỉ lệ giữa giao và hợp, hay nói cách khác là tìm tỉ lệ số phần tử giống nhau trên tổng số phần tử của 2 tập hợp. Do vậy, Jaccard score càng cao thì số lượng phần tử giống nhau càng cao hoặc tổng số phần tử khác nhau càng thấp, hai tập hợp càng giống nhau.
- Jaccard score được sử dụng trong bài toán này để tính toán độ tương đồng giữa tập label gốc và tập label được dự đoán.

3.3.4. Hamming Loss

- Hamming Loss là tỉ lệ dự đoán sai của từng label trên một tập hợp kết quả dự đoán. Dễ hiểu hơn, cho một vector **y** các nhãn thực và một vector **y'** các nhãn đã dự đoán của một mẫu, Hamming Loss trên mẫu đó chính là tỉ lệ label sai trên số lượng label

Và với một bộ dữ liệu với n mẫu và L nhãn, Hamming Loss được tính với công thức như sau:

$$\text{Hamming Loss} = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L}\mathbb{I}\left(y_{ij} \neq \hat{y}_{ij}\right)$$

- Điểm đặc biệt của Hamming Loss so với các chỉ số khác là sự nhạy cảm với từng nhãn dữ liệu hơn là với độ chính xác của cả tập hợp nhãn dữ liệu. Ngoài ra, Hamming Loss cũng là một chỉ số quan trọng trong bài toán multilabel classification nhờ việc tính toán trực tiếp trên từng label mà không phải là độ chính xác của tập kết quả, giúp việc xem xét và cải thiện kết quả dễ dàng hơn[5].

3.4. Các phương pháp tiền xử lý dữ liệu

3.4.1. Phương pháp tiền xử lý dữ liệu

Dữ liệu thô dạng thường tồn tại nhiễu và thiếu tính cấu trúc, tồn tại nhiều các từ viết tắt hay các thông tin không liên quan, việc xử lý dữ liệu giúp cho:

- **Cải thiện chất lượng dữ liệu:** Loại bỏ các thông tin không liên quan, để chắc chắn rằng dữ liệu đưa vào quá trình huấn luyện được sạch và nhất quán.
- **Nâng cao hiệu suất mô hình:** Văn bản được xử lý tốt có thể giúp mô hình trích xuất đặc trưng dễ hơn, cải thiện hiệu suất của các mô hình xử lý ngôn ngữ tự nhiên.
- **Giảm độ phức tạp:** Đơn giản hóa văn bản có thể giảm độ phức tạp tính toán và giúp mô hình hiệu suất hơn.

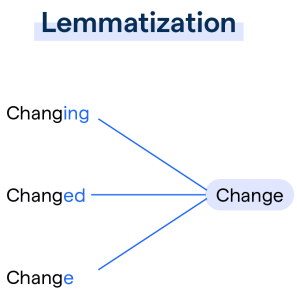
Trong xử lý ngôn ngữ tự nhiên, có nhiều phương pháp để xử lý văn bản, sau đây sẽ giới thiệu các phương pháp xử lý mà nhóm áp dụng. Các phương pháp mà nhóm chọn để thực hiện đã được lựa chọn để phù hợp với bộ dữ liệu mà nhóm đang thực hiện:

- **Loại bỏ các ký tự HTML:** Các ký tự HTML thường xuất hiện trong ngữ liệu do quá trình crawl từ các trang có thể lẫn các kí tự không cần thiết, đây là nhiễu cho dữ liệu. Các kí tự thường xuất hiện như , <html>, <a>,....
- **Loại bỏ các liên kết web:** Một số liên kết có thể xuất hiện trong mô tả do quá trình thu thập dữ liệu, thông thường, đây là lỗi của việc thu thập, loại bỏ các liên kết không liên quan có thể giúp dữ liệu giảm nhiễu.
- **Thay đổi từ viết gọn, viết tắt thành từ hoàn chỉnh:** Trong tiếng Anh thường có sự xuất hiện của các từ viết gọn như ‘ain’t’, ‘I’m’, ‘she’s’, ‘he’s’. Trong quá trình tách từ có thể khiến đầu ra không như mong muốn.

    + **Ví dụ:** Trong TF-IDF Vectorizer, quá trình Tokenizer của các câu:
        - "I am a cat" → ["I", "am", ""a", "cat"] → Như mong muốn.
        - "I'm a cat" → ["I", "m", "a", "cat"] → Từ "am" biến thành m.

- **Chuyển tất cả về ký tự in thường:** Việc này giúp đồng nhất các ký tự hoa và thường, làm tăng quá trình nhận dạng, thông thường được áp dụng trong các bài toán phân loại, do kết quả đầu ra chỉ cần biết thuộc lớp nào.
- **Loại bỏ chữ số và dấu câu:** Các số và dấu câu thường không mang ý nghĩa trong việc xác định loại phim, chỉ một số trường hợp ngoại lệ như:
    + 666: Horror
    + 1000 soldiers: War
    + Year 2049: Science Fiction

  Tuy nhiên, các trường hợp này xuất hiện rất hiếm xuất hiện nên nhóm quyết định xóa bỏ các chữ số và dấu câu.
- **Loại bỏ các biểu tượng cảm xúc và các kí tự cảm xúc:** Tuy trong bộ dữ liệu này, số lượng các kí tự hay biểu tượng cảm xúc không quá nhiều, nhưng nhóm vẫn quyết định cài đặt để tối ưu.
- **Loại bỏ các stopword**: Các stopword là các từ ít khả năng sẽ giúp mô hình đưa ra được kết luận. Một số stopword phổ biến như: the, in, with, because,... Loại bỏ stopword giúp mô hình tập trung vào các từ quan trọng và hiểu tốt hơn về ngữ cảnh và ý nghĩa của văn bản.
- **Biến đổi về từ gốc (lemmatization):** Trong tiếng Anh hay nhiều tiếng khác có sự xuất hiện của bổ ngữ, các từ có thể được biểu diễn ở dạng khác nhau với từng thời điểm quá khứ, hiện tại và tương lai hoặc các dạng biểu diễn động từ. Việc chuyển về dạng gốc của từ có thể giúp mô hình tăng được xác suất xuất hiện của từ ở từng thể loại, giúp tăng mức độ nhận diện.

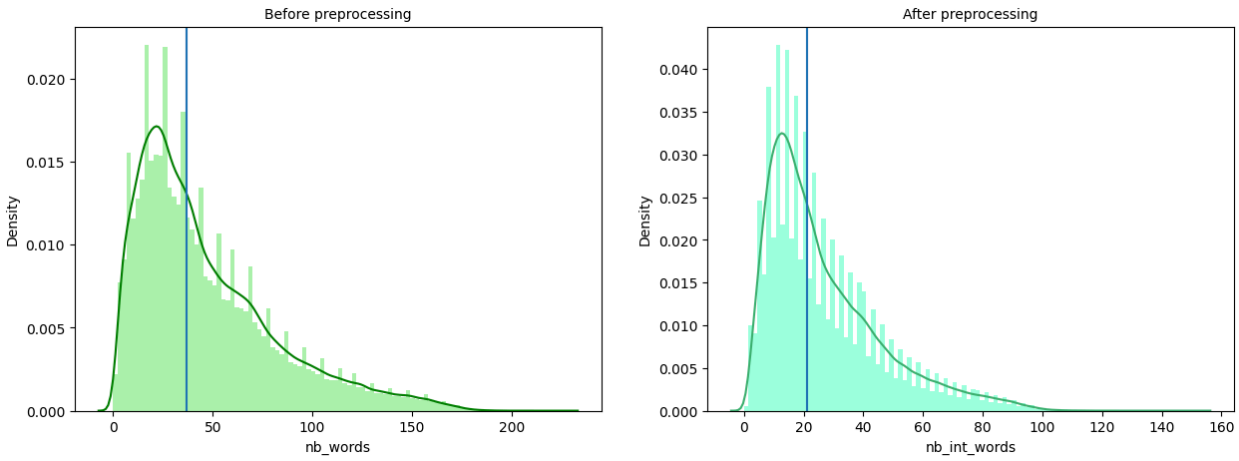


Hình 3.4.1 Mô tả cách Lemmatization thực hiện

Trước tiền xử lý	Sau tiền xử lý
Central Park is composed of 11 films that offer a visual sound and poetic journey through 11 cities crossed by the artist. Fascinated by the city space and urbanism Dominique Gonzalez-Foerster has been developing the concept of "tropical modernity" for several years starting from the cohabitation and the confrontation between architecture and vegetation. It is around this research that she designed "Central Park" 11 films : - Kyoto (1998) - Taipei (2000) - Buenos Aires (2003) - Los Glaciares (2003) - Hong Kong (2000) - Encore Tapei (2000) - White Sands (2003) - Brasilia (1998) - Paris (1999) - Shangai (2003) - Rio de Janeiro (2000)	central park compose film offer visual sound poetic journey city cross artist fascinate city space urbanism dominique gonzalezfoerster develop concept tropical modernity several year start cohabitation confrontation architecture vegetation around research design central park film kyoto taipei buenos aire los glaciares hong kong encore tapei white sand brasilia paris shangai rio de janeiro
“Set against the rattle of shopping carts and the white noise of L.A. traffic... “Disco’d” is an unvarnished moving look at the lives affected by the rising crisis of homelessness.” —Los Angeles Times	set rattle shopping cart white noise la traffic disco unvarnished move look life affect rise crisis homelessness los angeles time
http://KEXP.ORG presents black midi sharing an exclusive live performance with KEXP and talking to Larry Mizell Jr. host of The Afternoon Show. Recorded April 19 2021.	present black midi share exclusive live performance kexp talk larry mizell jr host afternoon show record april

Bảng 3.4.2. Bảng so sánh trước và sau khi tiền xử lý dữ liệu

Sau bước tiền xử lý dữ liệu, số lượng từ không cần thiết được lược bỏ, các liên kết web hay các kí hiệu, số không cần thiết cũng được xóa đi, giúp giảm số token mà mô hình nhận vào, cũng như giúp tối ưu hơn trong việc huấn luyện.



Hình 3.4.3. Số lượng từ trước và sau khi tiền xử lý

3.4.2. Phương pháp cắt mẫu

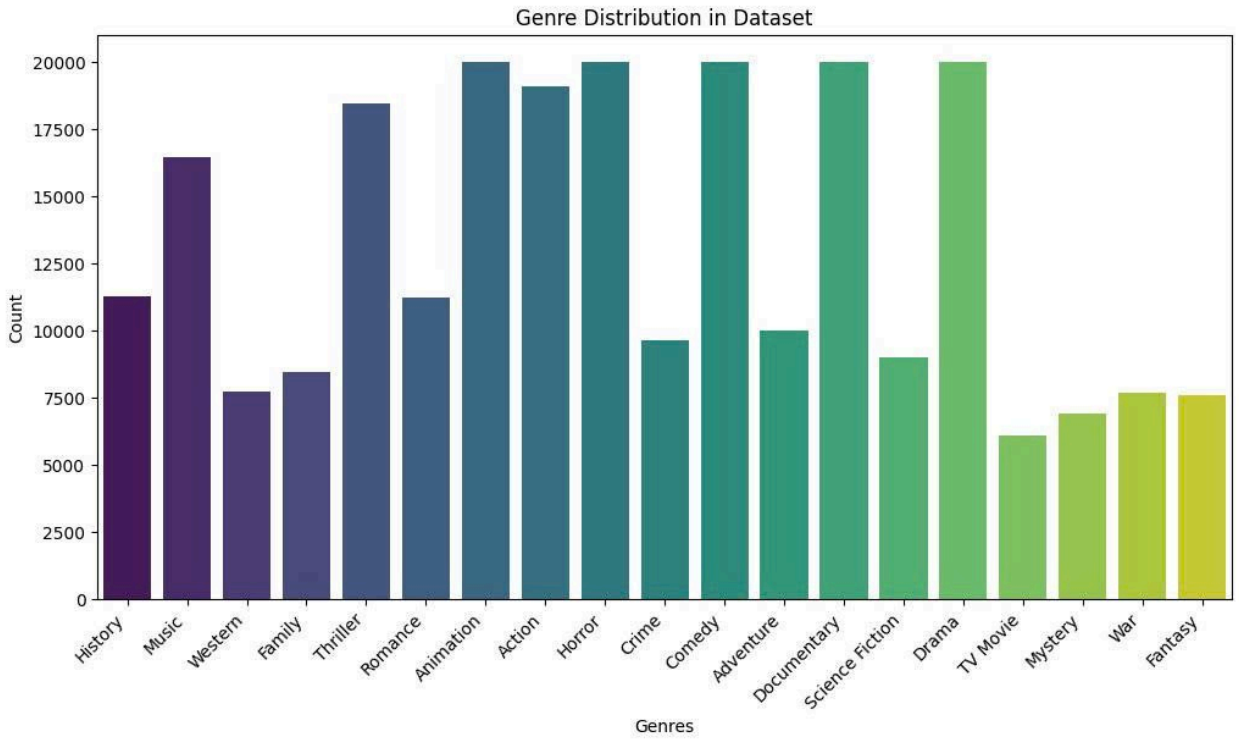
Sau khi EDA bộ dữ liệu đã được tiền xử lý ở trên, nhóm đề tài nhận thấy rằng sự chênh lệch giữa các nhãn (thể loại phim) là rất lớn.

Cụ thể, một số thể loại như Drama (153.869 mẫu), Documentary (98.237 mẫu), và Comedy (92.391 mẫu) chiếm tỷ lệ rất lớn trong toàn bộ tập dữ liệu. Ngược lại, các thể loại như History (11.268 mẫu), War (7.694 mẫu), và Western (7.701 mẫu) lại có số lượng xuất hiện rất thấp, chỉ bằng khoảng 1/10 so với các thể loại phổ biến.

Tình trạng mất cân bằng nhãn nghiêm trọng này ảnh hưởng lớn đối với bài toán phân loại đa nhãn (multi-label classification). Nếu không được xử lý hiệu quả, mô hình có xu hướng thiên vị các nhãn phổ biến hơn trong quá trình huấn luyện, dẫn đến việc bỏ qua các nhãn hiếm. Điều này có thể gây ra hiện tượng quá khớp (overfitting) trên các nhãn đa số và làm giảm khả năng khái quát hóa của mô hình, dẫn đến kết quả đánh giá không chính xác.

Để khắc phục vấn đề này, nhóm đề tài đã sử dụng phương pháp cắt mẫu. Quyết định này được đưa ra dựa trên hai yếu tố chính: kích thước khá lớn của bộ dữ liệu (435.706 mẫu) và sự chênh lệch đáng kể giữa các nhãn. Phương pháp cắt mẫu giúp điều chỉnh phân phối nhãn, giảm thiểu sự chênh lệch giữa các thể loại phim phổ biến và hiếm, từ đó cải thiện hiệu suất của mô hình phân loại đa nhãn.

Bắt đầu quá trình, nhóm xáo trộn dữ liệu để đảm bảo rằng các mẫu được phân bố ngẫu nhiên. Sau đó, nhóm thực hiện duyệt qua từng nhãn để tính tần số xuất hiện, giữ lại những mẫu chứa nhãn hiếm xuất hiện (dưới 10000 mẫu). Tiếp tục duyệt qua từng mẫu để thu những nhãn khác (tối đa là 20000 mẫu).



Hình 3.4.2.1. Phân phối của các lớp sau khi cắt giảm dữ liệu

Sau khi thực hiện phương pháp cắt mẫu, bộ dữ liệu đã được điều chỉnh xuống còn 157162 mẫu. Sự chênh lệch giữa các nhãn cũng đã giảm thiểu đáng kể. Các nhãn phổ biến như Drama và Comedy chỉ còn 20000 mẫu mỗi nhãn. Các nhãn hiếm như War, History và Western được giữ nguyên.

3.5. Transformers và các mô hình BERT

3.5.1. Kiến trúc Transformer

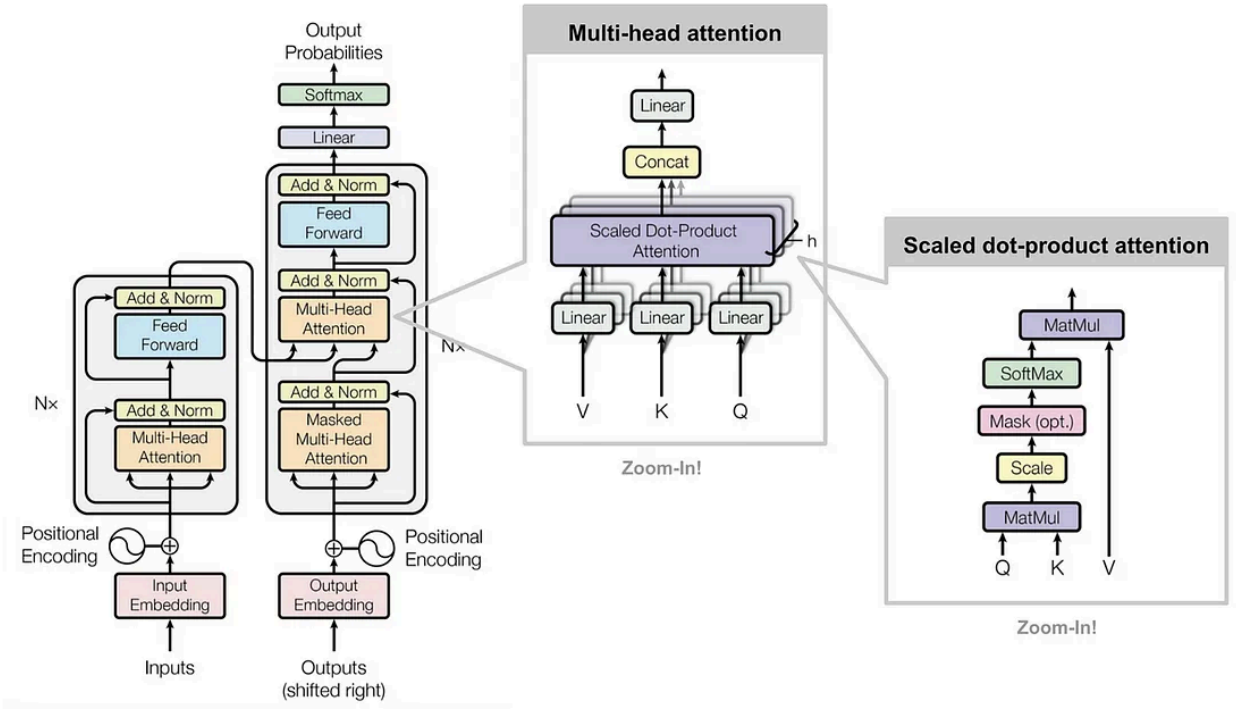
Transformer được sinh ra để giải quyết vấn đề gradient vanishing ở các mô hình họ RNN trước đó. Với hai khối chính là encoder và decoder:

+ Khối encoder: Mục tiêu chính vẫn là sinh ra một vector đại diện cho mỗi từ.
+ Khối decoder: Dùng để sinh từng từ trong đầu ra, dựa trên thông tin từng encoder và từ các từ đã sinh trước đó.

Các thành phần quan trọng trong kiến trúc Transformer:

- **Multi-Head Attention:** Thay vì có một attention duy nhất, mô hình xử lý song song, mỗi head học cách chú ý khác nhau, sau đó kết hợp lại với nhau.

- **Masked Multi-Head Attention:** Trong khối decoder, sử dụng phương pháp masking để không cho các mô hình nhìn thấy từ phía sau, việc này giúp mô hình sinh từ tuần tự theo đúng hướng
- **Positional Encoding:** Transformer không có cơ chế tuần tự như RNN nên cần khối này để mã hóa vị trí, dùng hàm sin//cos để thêm thông tin vị trí vào vector đầu vào. Tuy có hai khối là encoder và decoder, một các mô hình nổi tiếng hiện nay chỉ sử dụng một trong hai khối tùy vào cách ứng dụng của mô hình
    + **BERT:** Chỉ sử dụng khối encoder để áp dụng vào bài toán phân loại văn bản, trích xuất thông tin hay trả lời câu hỏi.
    + **GPT:** Chỉ sử dụng khối Decoder, áp dụng vào bài toán tạo sinh.



Hình 3.5.1.1 Kiến trúc Transformer

3.5.2. Các mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) là một kiến framework học máy mã nguồn mở được thiết kế cho xử lý ngôn ngữ tự nhiên. Bắt nguồn từ năm 2018, framework này được giới thiệu bởi Google AI Language. BERT tận dụng kiến trúc Transformer để hiểu vào tạo sinh một ngôn ngữ giống con người. Chỉ sử dụng khối encoder, BERT xử lý tốt các bài toán thiên về việc hiểu đầu vào hơn là bài toán tạo sinh câu[6].

Để giải quyết bài toán phân loại phim, nhóm quyết định chọn các mô hình được tiền huấn luyện sau:

- **BERT-base-uncased:** Là mô hình sử dụng 12 lớp Transformer, 768 hidden units và 12 attention heads, uncased nghĩa là không phân biệt chữ hoa và chữ thường, được huấn luyện trên bộ BookCopus và English Wikipedia. Lý do nhóm chọn mô hình này bởi vì các phim thường được lấy cảm hứng trên các sách hay tiểu thuyết, việc được huấn luyện trên BookCorpus có thể giúp mô hình dễ dàng hơn trong việc hiểu các mô tả có tính tự sự hay giàu ngữ cảnh.
- **BERT-base-cased:** Tương tự với mô hình trên, tuy nhiên, khác nhau ở case-insensitive, phiên bản này có phân biệt chữ hoa và chữ thường, được sử dụng nhiều trong các bài toán NER, khi chữ hoa có ý nghĩa quan trọng. Do dữ liệu đầu vào là đoạn tóm tắt của phim, một số tên nhân vật cũng đóng vai trò quan trọng, nhóm muốn thử huấn luyện trên mô hình này để so sánh xem có sự khác biệt nào về hiệu suất của mô hình so với uncased hay không.
- **DistilBERT-uncased:** Mô hình này được nén so với hai mô hình trên, tuy nhiên, theo giới thiệu, mô hình vẫn đạt được xấp xỉ 97% so với chưa nén, nhưng việc dự đoán vẫn nhanh hơn và số lượng tham số được giảm. Nhóm cũng muốn tối ưu thời gian dự đoán của mẫu dữ liệu, đáp ứng cho các công việc thời gian thực nên đã fine-tune trên mô hình này.

3.6. MultiOutput Logistic Regression

3.6.1. Logistic Regression

Mô hình Logistic Regression là mô hình học máy cơ bản, được sử dụng rộng rãi trong bài toán phân lớp, đặc biệt là phân lớp nhị phân. Logistic Regression hoạt động dựa trên nguyên tắc của hàm sigmoid - một hàm phi tuyến tự chuyển đầu vào của nó thành xác suất thuộc về một trong hai lớp nhị phân[7].

Hàm sigmoid được biểu diễn như sau:

$$S(z) = \frac{1}{1 + e^{-z}}$$

Đầu vào của hàm sigmoid là một giá trị z bất kỳ và trả về đầu ra là một giá trị nằm trong khoảng [0;1]. Khi áp dụng mô hình Logistic Regression với đầu vào là một ma trận dữ liệu X và trọng số w, thì ta được z = wX.

Việc huấn luyện mô hình là tìm ra bộ trọng số w sao cho đầu ra dự đoán của hàm sigmoid là giống với thực tế nhất. Để làm được thì chúng ta cần phải có hàm mất mát (loss function) để đánh giá hiệu năng của mô hình. Giá trị hàm mất mát càng nhỏ, mô hình dự đoán càng tốt. Trong bài toán dùng Logistic Regression, ta thường dùng hàm mất mát Cross-Entropy:

$$L(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

Trong quá trình huấn luyện, chúng ta tìm cách cập nhật bộ trọng số w sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự đoán tốt nhất.

3.6.2. MultiOutputClassifier

Trong bài toán phân loại đa nhãn (multi-label classification), mỗi mẫu dữ liệu có thể được gắn với nhiều nhãn khác nhau. Do đó, không thể áp dụng trực tiếp các mô hình phân lớp thông thường như Logistic Regression vào toàn bộ đầu ra nhiều nhãn.

Để giải quyết vấn đề này, nhóm đề tài đã sử dụng cấu trúc MultiOutputClassifier từ thư viện scikit-learn. Đây là một wrapper cho phép mở rộng bất kỳ mô hình phân loại nhị phân thành một mô hình đa nhãn, bằng cách huấn luyện một mô hình độc lập cho từng cột nhãn trong tập dữ liệu đầu ra.

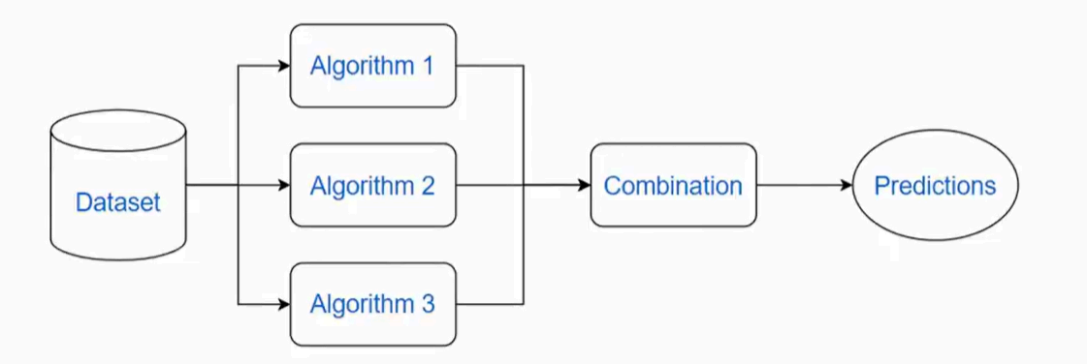
Nguyên lý hoạt động: Cho một tập dữ liệu gồm đầu vào là X và đầu ra đa nhãn với Y thuộc {0, 1}n x k với n là số mẫu và k là số nhãn. MultiOutputClassifier sẽ:

- Huấn luyện k mô hình phân loại nhị phân (ở đây sử dụng Logistic Regression).
- Mỗi mô hình sẽ dự đoán cho một nhãn (mỗi cột của Y).

Trong đề tài này, MultiOutputClassifier được kết hợp với LogisticRegression để tạo thành một pipeline phân loại cơ bản, dùng làm mô hình baseline nhằm so sánh với các mô hình học sâu như BERT. Sau khi huấn luyện, mô hình được lưu bằng joblib để có thể tái sử dụng trong giai đoạn dự đoán và đánh giá.

3.7. Ensemble

Ensemble là phương pháp sử dụng nhiều mô hình trong dự đoán thay vì một mô hình. Mỗi mô hình có thể mạnh ở một điểm nào đó, khi đặt kết quả của chúng chung, ta nhận được kết quả đúng hơn[8].



Hình 3.7.1 Mô tả Ensemble

Có nhiều loại Ensemble khác nhau như: Voting ensemble, Correlation, Averaging, Ranking averaging và Blending. Ở đây, nhóm sẽ tập trung vào Voting Ensembling.

Giả sử, có hai mô hình dự đoán trên cùng một mẫu, một mô hình cho ra kết quả là 0, một mô hình cho ra 1. Để giải quyết vấn đề trên, ta chọn thêm một mô hình nữa. Nhãn nào có số lượng dự đoán nhiều nhất sẽ là kết quả cuối cùng. Để tránh việc số lượng nhãn dự đoán bằng nhau, người ta thường chọn số lượng mô hình là số lẻ.

Chương 4. CÀI ĐẶT MÔ HÌNH VÀ THỰC NGHIỆM

4.1. Cài đặt hệ thống

Đối với nhiệm vụ xử lý dữ liệu, nhóm sử dụng các thư viện cơ bản như: nltk, re, string, emot, contraction. Do dữ liệu đầu vào khá lớn, để tải được dữ liệu, nhóm sử dụng polars thay vì pandas để tối ưu tốc độ xử lý.

Việc fine-tune được nhóm thực hiện trên nhiều nền tảng, bao gồm Google Colab, Kaggle. Các GPU được sử dụng là T4, T4 x2, P100. Đối với Logistic Regression, nhóm thực hiện fit mô hình trên TPU.

4.2. Cài đặt các phương pháp tiền xử lý dữ liệu

4.2.1. Phương pháp tiền xử lý dữ liệu

4.2.1.1. Tải xuống và đọc dữ liệu

Bước đầu, thực hiện tải dataset từ Hugging Face. Sau đó, loại bỏ các cột không cần thiết, loại bỏ các mẫu là NULL trong cột genres và overview. Ở đây chưa thể sử dụng df.drop_nulls() bởi vì có thể tồn tại giá trị NULL trong các cột không quan trọng.

```
df = pl.read_csv('hf://datasets/wykonos/movies/movies_dataset.csv')
df = df.select(df.columns[:5])
```

```
    df = df.filter(~(pl.col('genres').is_null() |
pl.col('overview').is_null()))
```

4.2.1.2. Loại bỏ các kí tự HTML

Các kí tự HTML như được giới thiệu ở phần trước đó, sẽ có dạng <tên thẻ>. Nhóm sử dụng thư viện regex để xử lý, nếu như đầu vào khớp với định dạng này, thì xóa bỏ.

```
import re, string

def remove_html(text):
    html_pattern = re.compile('<.*?>')
    return html_pattern.sub(r' ', text)
```

4.2.1.3. Loại bỏ các liên kết web

Tương tự, các liên kết web sẽ có dạng http hoặc https hoặc không có, sau đó là "://", "www". Nếu văn bản đầu vào khớp với định dạng này thì xóa.

```
def remove_urls(text):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub(r' ', text)
```

4.2.1.4. Thay đổi các từ viết gọn, viết tắt thành từ hoàn chỉnh

Ở đây, nhóm sử dụng thư viện contractions, gọi đến phương thức fix để biến từ gọn, từ viết tắt thành từ hoàn chỉnh.

```
import contractions

def expand_contractions(text):
   expanded_text = contractions.fix(text)
   return expanded_text
```

4.2.1.5. Loại bỏ chữ số và dấu câu

Nếu như văn bản không thuộc các kí tự từ a → z, A → Z, 0 → 9 , ' ' thì loại bỏ.

```
import contractions

def remove_special_characters(text):
    text = re.sub('[^a-zA-z0-9\s]', ' ', text)
    return text
```

4.2.1.6. Loại bỏ các biểu tượng cảm xúc và ký tự cảm xúc

Đối với nhiệm vụ này, nhóm sử dụng thư viện emot, thực hiện lấy ra các ký tự cảm xúc hay biểu tượng cảm xúc trong thư viện, gán vào một từ điển, nếu như văn bản khớp với các biểu tượng cảm xúc trong văn bản đó thì xóa, hàm này cũng hỗ trợ việc chuyển kí tự cảm xúc về đúng ý nghĩa của nó, tuy nhiên, do bài toán của nhóm thực hiện không có nhiều biểu tượng cảm xúc nên đã quyết định xóa đi.

```
import emot

def handle_emoticons(text, remove_emoticon=True):
    emot_obj = emot.emot()
    dict_emoticons = dict(zip(emot_obj.emoticons(text)['value'],
emot_obj.emoticons(text)['mean']))
    res_emoticons =  dict(sorted(dict_emoticons.items(), key = lambda
kv:len(kv[1]), reverse=True))
    for emoticon, mean in res_emoticons.items():
        if remove_emoticon:
            text = text.replace(emoticon, " ")
        else:
            text = text.replace(emoticon, mean)
    return text
```

4.2.1.7. Loại bỏ các từ dừng

Nhóm sử dụng bộ từ dừng có trong stopwords của nltk. Định nghĩa một set STOPWORDS: nếu như các từ được tách ra khớp với các từ trong STOPWORDS thì xóa bỏ. Ở đây sử dụng str(text).split() thay vì replace để tránh trường hợp các từ như ‘a’, sử dụng replace sẽ bị thay đổi không như mong muốn.

Ví dụ nếu dùng replace: remove_stopwords(‘apple’) → ‘pple’ → không mong muốn.

```
from nltk.corpus import stopwords

def remove_stopwords(text):
    """custom function to remove the stopwords"""
    STOPWORDS = set(stopwords.words('english'))
    return " ".join([word for word in str(text).split() if word not in
STOPWORDS])
```

4.2.1.8. Biến đổi về từ gốc (lemmatization)

Nhóm sử dụng hàm WordNetLemmatizer và bộ ngữ liệu wordnet từ thư viện nltk. Đầu tiên, chương trình định nghĩa một wordnet_map để ánh xạ ký hiệu POS (như 'N', 'V', 'J', 'R') sang các loại từ tương ứng trong WordNet.

Sau đó, sử dụng nltk.pos_tag để gán nhãn từ loại cho từng từ trong câu (sử dụng str(text).split() để tách từ). Từng từ sau đó sẽ được lemmatize bằng cách dùng lemmatizer.lemmatize() kết hợp với từ loại tương ứng. Nếu không xác định được POS rõ ràng, hàm sẽ mặc định dùng động từ (wordnet.VERB) để lemmatize.

Cuối cùng, kết quả được ghép lại thành chuỗi đầu ra đã được chuẩn hoá từ vựng.

Ví dụ: lemmatize_words("Cats are running quickly") → "cat be run quickly"

```
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet

def lemmatize_words(text):
    wordnet_map = {"N":wordnet.NOUN, "V":wordnet.VERB, "J":wordnet.ADJ,
"R":wordnet.ADV}
    lemmatizer = WordNetLemmatizer()
    pos_tagged_text = nltk.pos_tag(text.split())
    return " ".join([lemmatizer.lemmatize(word, wordnet_map.get(pos[0],
wordnet.VERB)) for word, pos in pos_tagged_text])
```

4.2.1.9. Tổng hợp các hàm tiền xử lý

Để gọi thực hiện tiền xử lý dữ liệu hàng loạt, nhóm tổng hợp các hàm lại thành một hàm duy nhất, giúp dễ dàng trong việc gọi một lần.

```
def preprocess_text(text):
    text = remove_html(text)
    text = remove_urls(text)
    text = expand_contractions(text)
    text = text.lower()
    text = remove_punc_and_num(text)
    text = remove_special_characters(text)
    text = handle_emoticons(text)
    text = remove_stopwords(text)
    text = lemmatize_words(text)
    return text
```

4.2.1.10. Gọi hàm preprocess

Gọi phương thức map_elements đến hàm preprocess để áp dụng tất cả các hàm tiền xử lý và dữ liệu.

```
    df_preprocessed = df.with_columns(
    pl.col('overview').map_elements(preprocess_text,
return_dtype=str).alias('overview')
    )
```

4.2.1.11. Áp dụng One-Hot-Encoding

Khởi tạo một set để lưu trữ các thể loại phim, duyệt qua các mẫu trong cột genres để tìm các thể loại phim khác nhau và thêm vào set distinct_genres. Duyệt qua cột genres, nếu như trong mẫu duyệt qua có tồn tại một trong những thể loại thì đánh dấu là 1, ngược lại, đánh dấu là 0.

```
    distinct_genres = set()
    for genre in df['genres']:
        genre = str.split(genre, '-')
        distinct_genres.update(genre)

    for genre in distinct_genres:
        df = df.with_columns(
            pl.when(pl.col("genres").str.contains(genre))
            .then(1)
            .otherwise(0)
            .alias(genre)
        )
```

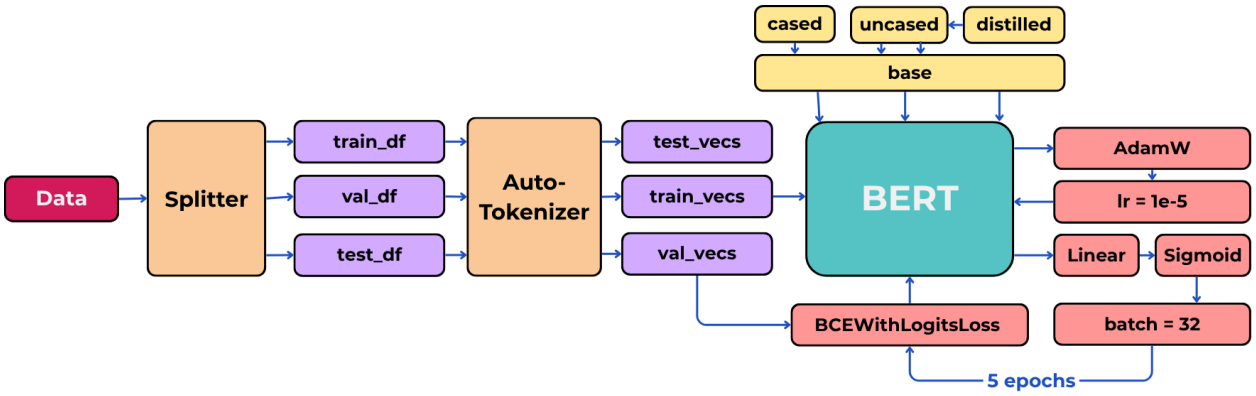
4.2.2. Phương pháp cắt mẫu

Xáo trộn bộ dữ liệu để đảm bảo tính ngẫu nhiên rồi duyệt lần lượt qua toàn bộ nhãn để tính tần số xuất hiện của một nhãn. Giữ lại nhãn cần bảo tồn bằng cách truyền tham số qua preserve_labels. Sau đó, kiểm tra từng nhãn rồi kiểm tra và thêm mẫu vào sao cho số lượng mẫu nhỏ hơn hoặc bằng max_per_genre.

```
def trim_dataset(df_train, max_per_genre=20000, preserve_labels = []):

  genre_cols = df_train.columns[1:]
```

```
  genre_counter = {genre: 0 for genre in genre_cols}

  df_shuffled = df_train.sample(fraction=1.0, shuffle=True, seed=42)
  kept_rows = []

  for row in df_shuffled.iter_rows(named=True):
      genres = [genre for genre in genre_cols if row[genre] == 1]
      if any(label in preserve_labels for label in genres):
          kept_rows.append(row)
          for genre in genres:
              genre_counter[genre] += 1

  for row in df_shuffled.iter_rows(named=True):
      genres = [genre for genre in genre_cols if row[genre] == 1]
      if row in kept_rows:
          continue

      if any(label in preserve_labels for label in genres):
          continue

      if all(genre_counter[genre] < max_per_genre for genre in genres):
          kept_rows.append(row)
          for genre in genres:
              genre_counter[genre] += 1

  df_trimmed = pl.DataFrame(kept_rows)
  return df_trimmed

df_trimmed = trim_dataset(df_train, max_per_genre = 20000,
preserve_labels=['War', 'History', 'Western'])
```

4.3. Cài đặt cho các mô hình BERT



Hình 4.3.1 Sơ đồ quá trình huấn luyện

4.3.1. Thiết lập các tham số chuẩn bị cho quá trình huấn luyện

Các hằng số được sử dụng thường xuyên:

- max_length (Số từ trong một câu của câu dài nhất): 180
- num_labels (Số lượng nhãn, tương ứng với số thể loại phim): 19

4.3.2. Chuẩn bị dữ liệu

Đọc dữ liệu từ tệp đã được xử lý, sau đó, nhóm loại bỏ đi các cột không cần thiết bao gồm "original_language", "id", "genres", "title". Nhóm chia dữ liệu theo tỉ lệ 70:20:10 tương ứng cho tập train:val:test, sử dụng hàm train_test_split trong thư viện scikit-learn để thực hiện việc chia các tập dữ liệu. Do quá trình huấn luyện có thể được thực hiện trên nhiều biến thể của bộ dữ liệu, nhóm đồng bộ thứ tự các lớp bằng một tệp json như sau.

```
{
    "0": "Comedy",
    "1": "Fantasy",
    "2": "Crime",
    "3": "Family",
    "4": "Horror",
    "5": "TV Movie",
    "6": "Action",
    "7": "Animation",
    "8": "War",
    "9": "Documentary",
    "10": "Western",
    "11": "History",
```

```
    "12": "Thriller",
    "13": "Mystery",
    "14": "Music",
    "15": "Romance",
    "16": "Drama",
    "17": "Adventure",
    "18": "Science Fiction"
}
```

Do trong các lần huấn luyện trước đó không thực hiện đồng bộ thứ tự của các lớp, nhóm nhận thấy việc sai lệch trong quá trình dự đoán, nên trong source code của nhóm có 3 tệp json khác nhau để giải mã cho thứ tự của từng mô hình. Tên của các loại phim được lưu vào target_list.

4.3.3. Cài đặt lớp token hóa

Để thực hiện tokenize cho các mô hình BERT, nhóm sử dụng AutoTokenizer, gọi đến phương thức from_pretrained có trong thư viện transformer, nhận vào đối số là trên của mô hình được tiền huấn luyện.

- Ta phải định nghĩa một lớp là CustomDataset nhận vào:
  + DataFrame (polars), chứa thông tin phim.
  + Tokenizer: Là AutoTokenizer được định nghĩa ở trên.
  + max_len: Là hằng số được khởi tạo trước đó (180).
  + target_list: Là danh sách các nhãn cần được dữ đoán, được khai báo trước đó.
- Sau đó, lớp này sẽ khởi tạo:
  + tokenizer: Là đối số truyền vào.
  + df: Là đối số truyền vào.
  + title: Là trường overview trong df.
  + targets: Là dữ liệu trong df của các trường trong target_list.
  + max_len: Là đối số truyền vào.
- Hàm __getitem__ nhận vào là index, sau đó:
  + Thực hiện xử lý bằng việc đổi title ở vị trí index thành kiểu string để chắc chắn.
  + Loại bỏ các khoảng trắng thừa trong title.
  + Sau đó, gọi phương thức encode_plus trong tokenizer với các đối số:
    - title

- None: Tham số thứ 2 dành cho văn bản thứ hai (cặp câu), bỏ qua vì đây là bài toán 1 câu.
- add_special_token=True, thêm token đặc biệt như [CLS], [SEP] – bắt buộc với BERT.
- max_length: self.max_len, định nghĩa độ dài tối đa của câu.
- padding='max_length', nếu như câu không đủ chiều dài tối đa thì thêm [PAD] cho đủ.
- return_token_type_ids = True, giúp trả về loại token, phân biệt câu 1/câu 2 trong BERT (luôn là 0 nếu chỉ 1 câu).
- truncation=True, sẽ cắt câu nếu như chiều dài câu vượt quá token_length.
- return_attention_mask=True, trả về attention mask (1: token thật, 0: padding).
- return_tensors='pt', trả kết quả dưới dạng Tensor PyTorch (thay vì list thông thường).

+ Hàm này sẽ trả về:

- Trường input_ids, attention_mask, token_type_ids trong inputs, được flatten, vì encode_plus trả về tensor có shape [1, seq_len], nên cần flatten() để biến về [seq_len] cho tương thích với mô hình.
- targets trả về dưới dạng vectors nhãn multilabel, title sẽ trả về văn bản gốc.

Thực hiện token hóa: Để thực hiện token hóa, ta gọi lớp CustomDataset đã được định nghĩa, truyền vào dataframe cần được token hóa, tokenizer là AutoTokenizer, max_len và target_list được nhắc đến trước đó. Dưới đây là mã nguồn ví dụ cho việc token hóa trên tập train.

```
tokenizer = AutoTokenizer.from_pretrained(pretrain_model_name)
train_dataset = CustomDataset(train_df, tokenizer, MAX_LEN, target_list)
train_data_loader = torch.utils.data.DataLoader(train_dataset,
    batch_size=TRAIN_BATCH_SIZE,
    shuffle=True,
    num_workers=0
    )
```

Ta gọi torch.utils.data.DataLoader để chia dữ liệu thành các batch, đồng thời, xáo trộn dữ liệu. Thực hiện tương tự với các tập val và test, ta có được train_data_loader, val_data_loader và test_data_loader.

4.3.4. Cài đặt các tham số huấn luyện

Các siêu tham số được áp dụng cho họ BERT được cài đặt đồng nhất như sau:

- Optimizer: AdamW,
- Learning rate: 1e-5,
- Epochs: 5,
- Batch size: 32,
- Activation function: Sigmoid,
- Loss function: BCEWithLogitsLoss,
- Final layer: Linear (768, 19), do có 19 thể loại phim.

4.3.5. Quá trình huấn luyện

Hàm train_model: Để thực hiện quá trình huấn luyện mô hình, ta cài đặt một hàm train_model, nhận vào là các training_loader, pretrained model, và optimizer. Một epoch sẽ được thực hiện theo các bước sau:

- Trước tiên sẽ duyệt qua các batch, lấy các thông tin như id, mask, token_type_ids và targets.
- Ở bước forward, truyền vào cho pretrained model các thông tin đã lấy được, bao gồm ids, mask, token_type_ids, kết quả mô hình tính toán được sẽ được lưu vào biến outputs.
    + Hàm loss_fn được định nghĩa sẵn với BCEWithLogitsLoss trong torch.nn sẽ tính loss giữa đầu ra và nhãn thực tế trong targets.
    + Áp đầu ra tính được của mô hình vào hàm sigmoid để ra giá trị nhị phân, ngưỡng được chọn là 0.5 như đã được nhắc đến.
    + Với mỗi nhãn dự đoán đúng sẽ được cộng vào biến correct_predictions.
    + Biến num_samples được dữ để tính tổng các nhãn (ví dụ: 32 sample × 19 labels = 608).
- Ở bước backward, trước tiên ta đặt lại gradient. Để tính đạo hàm, ta gọi loss.backward(). Và để chống tình trạng exploding gradient, ta dùng thêm tham số nn.utils.clip_grad_norm_. Gọi optimizer.step() để cập nhật trọng số optimizer.

```
for batch_idx, data in loop:
    ids = data['input_ids'].to(device, dtype = torch.long)
    mask = data['attention_mask'].to(device, dtype = torch.long)
    token_type_ids = data['token_type_ids'].to(device, dtype = torch.long)
    targets = data['targets'].to(device, dtype = torch.float)
```

```
        # forward
        outputs = model(ids, mask, token_type_ids)
        loss = loss_fn(outputs, targets)
        losses.append(loss.item())
        # training accuracy, apply sigmoid, round (apply thresh 0.5)
        outputs = torch.sigmoid(outputs).cpu().detach().numpy().round()
        targets = targets.cpu().detach().numpy()
        correct_predictions += np.sum(outputs==targets)
        num_samples += targets.size   # total num of elements in the 2D array

        # backward
        optimizer.zero_grad()
        loss.backward()
        nn.utils.clip_grad_norm_(model.parameters(), max_norm=1.0)
        # grad descent step
        optimizer.step()
```

Sau một epoch, hàm này sẽ trả về mô hình đã được huấn luyện (một epoch), accuracy của mô hình và trung bình loss ở thời điểm đó.

Hàm eval_model: Trong quá trình huấn luyện, ta cần đánh giá mô hình để biết được khi nào mô hình đã hội tụ, thực hiện dừng việc huấn luyện mô hình. Các tham số của hàm eval_model tương tự như hàm train_model, bổ sung thêm một vài tham số là val_loader, model và optimizer.

- Trước tiên, duyệt qua các batch và lấy ra các thông tin cần thiết, tương tự như việc huấn luyện.
- Truyền các thông tin lấy được vào mô hình và gán vào biến outputs.
- Thực hiện tính loss và gán vào biến loss, thêm biến loss vào mảng losses.
- Áp hàm sigmoid vào để chuyển output sang giá trị nhị phân.

```
 with torch.no_grad():
        for batch_idx, data in enumerate(validation_loader, 0):
            ids = data['input_ids'].to(device, dtype = torch.long)
            mask = data['attention_mask'].to(device, dtype = torch.long)
            token_type_ids = data['token_type_ids'].to(device, dtype =
torch.long)
            targets = data['targets'].to(device, dtype = torch.float)
            outputs = model(ids, mask, token_type_ids)
```

```
            loss = loss_fn(outputs, targets)
            losses.append(loss.item())

            # validation accuracy
            # add sigmoid, for the training sigmoid is in BCEWithLogitsLoss
            outputs = torch.sigmoid(outputs).cpu().detach().numpy().round()
            targets = targets.cpu().detach().numpy()
            correct_predictions += np.sum(outputs==targets)
            num_samples += targets.size   # total number of elements in the 2D
array

                # Subset accuracy (all labels correct per sample)
            subset_correct += np.sum(np.all(outputs == targets, axis=1))
        subset_acc = float(subset_correct) / num_samples
```

Kết quả trả về của hàm là accuracy, trung bình loss và subset accuracy

Bắt đầu huấn luyện: Cho một vòng lặp, được lặp với hằng số EPOCHS được nhắc đến trước đó, sau đó:

- Gọi đến hàm train_model, truyền vào train_data_loader, model, optimizer, giá trị trả về lưu vào các biến model, train_acc, train_loss.
- Thực hiện đánh giá mô hình bằng cách gọi hàm eval_model, truyền vào val_data_loader, model, optimizer, giá trị trả về lưu vào các biến val_acc, val_loss, val_subset_acc.
- Nếu như val_acc lớn hơn best_accuracy (được khởi tạo là 0) thì lưu mô hình vào đường dẫn được chỉ định.

```
    for epoch in range(1, EPOCHS+1):
        print(f'Epoch {epoch}/{EPOCHS}')
        model, train_acc, train_loss = train_model(train_data_loader, model,
optimizer)
        val_acc, val_loss, val_subset_acc = eval_model(val_data_loader, model,
optimizer)

        print(f'train_loss={train_loss:.4f}, val_loss={val_loss:.4f}
train_acc={train_acc:.4f}, val_acc={val_acc:.4f}')
```

```
        history['train_acc'].append(train_acc)
        history['train_loss'].append(train_loss)
        history['val_acc'].append(val_acc)
        history['val_loss'].append(val_loss)
        # save the best model
        if val_acc > best_accuracy:
            os.makedirs(save_dir, exist_ok=True)
            torch.save(model.state_dict(), os.path.join(save_dir,
"best_model.pt"))
            best_accuracy = val_acc
```

4.3.6. Hậu xử lý kết quả

Do có sự chênh lệch giữa các lớp và trong quá trình thực nghiệm, nhóm nhận ra một số lớp có confidence score thấp. Để giải quyết vấn đề này, nhóm kết hợp phương pháp lọc lại ngưỡng cho kết quả dự đoán: tăng ngưỡng confidence score đối với các lớp có độ chính xác cao, giảm đối với các lớp có chỉ số thấp. Dưới đây là ngưỡng mà nhóm đặt ra cho các lớp sau nhiều lần thực nghiệm:

Genre	Threshold	Genre	Threshold
Comedy	0.5	Documentary	0.6
Fantasy	0.3	Western	0.6
Crime	0.3	History	0.3
Family	0.3	Thriller	0.3
Horror	0.5	Mystery	0.3
TV Movie	0.3	Music	0.6
Action	0.3	Romance	0.3
Animation	0.6	Drama	0.5
War	0.3	Adventure	0.3
		Science Fiction	0.5

Bảng 4.3.6.1. Bảng ngưỡng confidence score cho từng nhãn

Các lớp có độ chính xác cao như Animation, Documentary, Western hay Music được cho ngưỡng là 0.6, các lớp khó hơn được chọn là 0.3. Tương tự đối với các lớp còn lại. Và để tránh trường hợp các lớp dự đoán đều thấp, nhóm quyết định sử dụng thêm phương pháp top k với k = 3, sau đó, sử dụng phép OR logic giữa hai phương pháp threshold và top k.

4.4. Miêu tả quá trình thực hiện trên BERT-base-uncased bằng một mẫu dữ liệu

Mẫu dữ liệu thô được chọn:

Central Park is composed of 11 films that offer a visual sound and poetic journey through 11 cities crossed by the artist. Fascinated by the city space and urbanism Dominique Gonzalez-Foerster has been developing the concept of "tropical modernity" for several years starting from the cohabitation and the confrontation between architecture and vegetation. It is around this research that she designed "Central Park" 11 films : - Kyoto (1998) - Taipei (2000) - Buenos Aires (2003) - Los Glaciares (2003) - Hong Kong (2000) - Encore Tapei (2000) - White Sands (2003) - Brasilia (1998) - Paris (1999) - Shangai (2003) - Rio de Janeiro (2000)

Áp dụng phương pháp tiền xử lý:

central park compose film offer visual sound poetic journey city cross artist fascinate city space urbanism dominique gonzalezfoerster develop concept tropical modernity several year start cohabitation confrontation architecture vegetation around research design central park film kyoto taipei buenos aire los glaciares hong kong encore tapei white sand brasilia paris shangai rio de janeiro

Cho qua hàm AutoTokenizer:

Mỗi từ tương ứng với ID trong vocabulary của tokenizer. Có một vài lưu ý về các token đặc biệt như:

- 101 = [CLS] token: Đánh dấu bắt đầu câu.
- 102 = [SEP] token: Đánh dấu kết thúc câu hoặc tách giữa hai cặp câu.
- Các số 0 là padding được thêm vào để cho đủ độ dài của câu.

```
tensor([[  101,  2430,  2380, 17202,  2143,  3749,  5107,  2614, 13805,  4990,
          2103,  2892,  3063,  6904, 11020, 14776,  2103,  2686,  3923,  2964,
         18165, 10121, 14876,  2545,  3334,  4503,  4145,  5133,  2715,  3012,
          2195,  2095,  2707,  2522, 25459, 18557, 13111,  4294, 10072,  2105,
```

```
       2470,  2640,  2430,  2380,  2143, 15008, 14004,  9204,  2250,  2063,
       3050,  1043,  2721,  7405,  6072,  4291,  4290, 19493,  6823,  2072,
       2317,  5472, 21133,  2401,  3000, 29382,  4886,  5673,  2139, 11497,
        102,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0,
          0,     0,     0,     0,     0,     0,     0,     0,     0,     0]])
```

Ngoài ra, khi cho vào AutoTokenizer, còn trả về hai tensor vectors là token_type_ids và attention_mask. Các giá trị trả về sẽ là 0 hoặc 1.

Cho qua mô hình được tiền huấn luyện

Sau khi chuyển các vector cần thiết bao gồm input_ids, attention_mask, token_type_ids vào mô hình nhận được kết quả đầu ra như sau:

```
tensor([[ 0.0237,  0.0836, -0.2832, -0.2055, -0.4075,  0.4432, -0.5060,  0.8169,
          0.2444,  0.1496,  0.0480,  0.2639, -0.1155, -0.2276,  0.2105,  0.0369,
         -0.0081,  0.0433, -0.0806]], grad_fn=<AddmmBackward0>)
```

Đầu ra là một tensor vector bao gồm 19 con số, tương ứng với 19 lớp

Áp dụng hàm sigmoid

Khi áp dụng hàm sigmoid vào kết quả từ mô hình, được đầu ra như sau:

```
array([[1., 1., 0., 0., 0., 1., 0., 1., 1., 1., 1., 1., 0., 0., 1., 1., 0., 1., 0.]], dtype=float32)
```

Tính loss:

Ở đây, ta có ground truth là một tensor như sau:

```
tensor([[0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```

Dựa vào kết quả đầu ra của mô hình, và ground truth, ta tính loss bằng hàm BCEWithLogitsLoss.

Nhắc lại, công thức của BCEWithLogitsLoss:

$$\text{BCEWithLogitsLoss}(x, y) = -[y \cdot \log(\sigma(x)) + (1 - y) \cdot \log(1 - \sigma(x))]$$

Ví dụ trên một vài lớp:

$$\sigma(x_0) = \frac{1}{1 + e^{-0.0237}} \approx 0.5059$$
$$\text{loss}_0 = -\log(1 - 0.5059) \approx -\log(0.4941) \approx 0.7054$$
$$\sigma(x_1) = \frac{1}{1 + e^{-0.0836}} \approx 0.5209$$
$$\text{loss}_1 = -\log(1 - 0.5209) \approx -\log(0.4791) \approx 0.7368$$
$$\sigma(x_2) = \frac{1}{1 + e^{0.2832}} \approx 0.4296$$
$$\text{loss}_2 = -\log(1 - 0.4296) \approx -\log(0.5704) \approx 0.5612$$

Từ các kết quả tính toán được, thực hiện lấy trung bình, ở đây có 19 lớp, sau khi lấy trung bình, ta được giá trị loss là **0.7102.**

Sau đó, mô hình sẽ tính gradient bằng cách gọi backward. Mô hình được cập nhật theo gradient tính được. Ở dây, weight và bias nhận được các giá trị lần lượt là **0.639215** và **0.034238.**

4.5. Cài đặt cho MultiOuput Logistic Regression

4.5.1. TF-IDF Vectorizer

Trong trích xuất thông tin, TF-IDF (term frequency - inverse document frequency) là độ đo mức độ quan trọng của một từ trong một tài liệu, ngữ liệu, bằng việc tính toán tần số xuất hiện của một từ.

Công thức:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Trong đó:

- TF: Tần số xuất hiện của từ trong document, được tính bằng cách lấy số lần xuất hiện của từ chia cho tổng số từ.
- IDF: Dùng để đánh giá mức độ quan trọng của một từ trong văn bản. Khi tính TF mức độ quan trọng của các từ là như nhau, tuy nhiên, trong văn bản thường xuất hiện các từ không quan trọng với tần suất cao, được tính bằng cách lấy log cơ số e của số văn bản trong tập D chia số văn bản chưa từ trong tập D.

Ví dụ cách tính TF-IDF:

Ở đây, ta sẽ tính TF-IDF của từ "Vietnamese" trong các mẫu sau:

Document	**TF**	**IDF**	**TF-IDF**
The story of four patriotic **Vietnamese** boys who join the North **Vietnamese** army in their fight to liberate South Vietnam.	$\frac{2}{20}$	$\frac{1}{30}$	$\frac{1}{3}$
A young **Vietnamese** girl must find her lost family after her city was destroyed by US's bombing campaign in 1972.	$\frac{1}{21}$	$\frac{1}{63}$	
An unlikely bond forms between an underground debt collector and a cai luong "**Vietnamese** opera" performer against the backdrop of Saigon in the 90s.	$\frac{1}{24}$	$\frac{1}{72}$	

Khi chỉ số TF-IDF càng cao thì mức độ ảnh hưởng của từ đến mô hình càng lớn.

4.5.2. Quá trình huấn luyện

Nhóm đề tài bắt đầu bằng việc xử lý dữ liệu, sử dụng bộ dữ liệu đầu vào đã được tiền xử lý ở bước trước (bao gồm chuẩn hóa văn bản, loại bỏ ký tự đặc biệt, lemmatization, v.v.). Để đảm bảo tính nhất quán và tránh lỗi khi huấn luyện mô hình, nhóm tiếp tục kiểm tra và loại bỏ các dòng có giá trị null ở cột mô tả nội dung phim (overview).

```
df = df.filter(~pl.col("overview").is_null())
overview = df["overview"].to_list()
```

Sử dụng các nhãn đã được one-hot encoding trước đó, bắt đầu từ Comedy. Việc chọn nhãn được thực hiện bằng cách lấy toàn bộ các cột từ vị trí Comedy cho đến cuối dataframe rồi chuyển về dạng numpy array.

```
label_cols = df.columns[df.columns.index("Comedy"):]
labels = df.select(label_cols).to_numpy()
```

Tiếp theo, nhóm tiến hành vector hóa dữ liệu văn bản từ cột overview bằng kỹ thuật TF-IDF (Term Frequency - Inverse Document Frequency). Việc vector hóa được thực hiện bằng hàm TfidfVectorizer() từ thư viện scikit-learn, với các tham số mặc định bao gồm:

- lowcase=True: tự động chuyển từ viết hoa về viết thường
- stop_words=None: không loại bỏ stop words (vì đã thực hiện ở preprocess)
- norm = "l2" : chuẩn hóa vector đầu ra theo chuẩn l2
- ngram_range=(1, 1): chỉ xét đơn từ (unigram)

```
tfidf = TfidfVectorizer()
vectors = tfidf.fit_transform(overview).toarray()
```

Sau khi đã vector hóa xong, nhóm tiến hành chia tập huấn luyện (train set) và tập kiểm thử (test set) theo tỉ lệ 8:2.

```
TEST_SIZE = 0.2
split_idx = int(TEST_SIZE * len(vectors))
Trainvecs = vectors[:split_idx]
Testvecs = vectors[split_idx:]
Trainclss = labels[:split_idx]
Testclss = labels[split_idx:]
```

Nhóm đề tài sử dụng một mô hình học máy cơ bản là Logistic Regression để thực hiện phân loại. Tuy nhiên, do bài toán đặt ra là phân lớp đa nhãn (multi-label classification) — tức mỗi mẫu có thể mang nhiều nhãn đồng thời — nhóm đã áp dụng kỹ thuật MultiOutputClassifier từ thư viện scikit-learn.

```
model = MultiOutputClassifier(LogisticRegression())
model.fit(Trainvecs, Trainclss)
```

Sau khi hoàn thành quá trình huấn luyện, nhóm tiến hành lưu lại:

- Trọng số mô hình vào logistic_regression_model.joblib
- Vectorizer TF-IDF đã được huấn luyện vào tệp tfidf_vectorizer.joblib

Việc lưu mô hình được thực hiện bằng thư viện joblib, cho phép tái sử dụng mô hình trong giai đoạn dự đoán hoặc đánh giá mà không cần huấn luyện lại từ đầu:

```
joblib.dump(model, "./weights/logistic_regression_model.joblib")
joblib.dump(tfidf, "./weights/tfidf_vectorizer.joblib")
```

4.6. Miêu tả quá trình thực hiện trên Logistic Regression bằng một mẫu dữ liệu

Giả sử bộ dữ liệu ba mẫu như sau:

The story of four patriotic Vietnamese boys who join the North Vietnamese army in their fight to liberate South Vietnam.

A young Vietnamese girl must find her lost family after her city was destroyed by US's bombing campaign in 1972.
An unlikely bond forms between an underground debt collector and a cai luong "Vietnamese opera" performer against the backdrop of Saigon in the 90s.

Chọn mẫu dữ liệu như sau, y = 1.

The story of four patriotic **Vietnamese** boys who join the North **Vietnamese** army in their fight to liberate South Vietnam.

Sau khi tiền xử lý dữ liệu:

story patriotic vietnamese boy join north vietnamese army fight liberate south vietnam

Vector hóa dữ liệu bằng TF-IDF thu được như sau:

```
array([0.        , 0.        , 0.29623539, 0.        , 0.        ,
       0.        , 0.29623539, 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.29623539,
       0.        , 0.        , 0.        , 0.29623539, 0.29623539,
       0.        , 0.        , 0.29623539, 0.        , 0.29623539,
       0.        , 0.        , 0.29623539, 0.29623539, 0.        ,
       0.        , 0.        , 0.29623539, 0.34992278, 0.        ])
```

Tính đầu ra của hàm sigmoid, với các trọng số khởi tạo ban đầu bằng 0 và learning rate = 0.01.

$$z = \mathbf{w}^\top \cdot \mathbf{x} = 0 \Rightarrow \hat{y} = \frac{1}{1 + e^{-0}} = 0.5$$

Tính toán hàm mất mát:

$$\mathcal{L}(0.5, 1) = -\left[1 \cdot \log(0.5) + (1 - 1) \cdot \log(1 - 0.5)\right] = -\log(0.5) \approx 0.693$$

Gradient của hàm mất mát theo w:

$$\mathbf{w}: \quad \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = (\hat{y} - y) \cdot \mathbf{x} = (0.5 - 1) \cdot \mathbf{x} = -0.5 \cdot \mathbf{x}$$

Cập nhật trọng số:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} - 0.01 \cdot (-0.5 \cdot \mathbf{x}) = 0.005 \cdot \mathbf{x}$$

Cập nhật bias:

$$b^{(t+1)} = b^{(t)} - \eta \cdot (\hat{y} - y) = 0 - 0.01 \cdot (-0.5) = 0.005$$

Sau một bước học mô phỏng, trọng số w được cập nhật thành như sau:

```
w = array([0.        , 0.        , 0.00148118, 0.        , 0.        ,
       0.        , 0.00148118, 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.00148118,
       0.        , 0.        , 0.        , 0.00148118, 0.00148118,
       0.        , 0.        , 0.00148118, 0.        , 0.00148118,
       0.        , 0.        , 0.00148118, 0.00148118, 0.        ,
       0.        , 0.        , 0.00148118, 0.00174961, 0.        ])
```

Áp dụng tương tự đối với 18 nhãn còn lại qua MultiOutputClassifier.

4.7. Kết quả thử nghiệm

Mục tiêu của của đánh giá kết quả thử nghiệm là để so sánh các mô hình trong nhiệm vụ phân loại đa nhãn (multi-label classification) thể loại phim dựa trên phần mô tả nội dung phim (overview). Các mô hình được đánh giá trên tập kiểm thử (test set) được chia tách từ bộ dữ liệu gốc theo tỉ lệ đã định sẵn và không xuất hiện trong tập huấn luyện (training set) và tập đánh giá (validation set).

Dữ liệu đầu vào là phần mô tả nội dung phim đã được tiền xử lý và token hóa theo từng tokenizer tương ứng. Kết quả đầu ra là một vector nhị phân là dạng one-hot encoding sau đó được ánh xạ ngược để thu về nhãn gốc.

Metrics đánh giá bao gồm các thông số: micro F1-score, Hamming loss và Jaccard accuracy như đã giới thiệu ở phần 3.2.

4.7.1. Mô phỏng kết quả dự đoán trên các mô hình và ensemble

Để trực quan hơn trong việc so sánh, nhóm lấy hai mẫu dữ liệu như sau:

Mẫu dữ liệu thô:

Porto (Portugal) beginning of the 80s. A film crew seeks to understand if there's really the figure of Chico Fininho immortalized in the popular song of the same name. A

bunch of people are interviewed taking us to Boavista - where we find him in a club. After a brief hesitation the interview happens and we're taken to everyday situations that can happen to any "freak" of the city...

Mẫu dữ liệu được tiền xử lý:

porto portugal beginning film crew seek understand really figure chico fininho immortalize popular song name bunch people interview take us boavista find club brief hesitation interview happen take everyday situation happen freak city

Ground truth: Drama, Music

Thực hiện inference trên các mô hình:

Mô hình	**Mẫu dữ liệu**	**Kết quả**
Bert-base-cased	Thô	TV Movie Comedy **Drama**
Bert-base-uncased	Được tiền xử lý	Documentary **Music** **Drama**
DistilBert	Được tiền xử lý	Documentary **Music** **Drama**
Ensemble	Thô, được tiền xử lý	Documentary **Drama** **Music**

4.7.2. Kết quả của các mô hình BERT trên tập kiểm thử

Dataset	**Model**	**Multi-label Classification**		
		micro-F1	**Hamming Loss**	**Jaccard Accuracy**
Raw + trimmed	Bert-base-cased	0.619079	0.053693	0.56525
Preprocessed	DistilBERT	0.62	**0.052731**	**0.5935**

Preprocessed + trimmed	Bert-base-uncased	**0.648779**	0.063554	0.589834

Dựa trên kết quả sau khi thực hiện đánh giá trên tập kiểm thử, nhóm nhận thấy rằng:

- Mô hình BERT-base-uncased có micro-F1 cao nhất (0.648779) thực hiện trên bộ dữ liệu đã được tiền xử lý và cắt ngắn. Điều này cho thấy việc loại bỏ thông tin dư thừa và giữ lại nội dung cốt lõi giúp mô hình cải thiện khả năng phân loại chính xác theo từng nhãn.
- Mô hình DistilBERT, mặc dù nhẹ hơn, vẫn thể hiện hiệu quả cao khi được huấn luyện trên dữ liệu đã tiền xử lý (preprocessed). Cụ thể:
  + Đạt Hamming Loss thấp nhất (0.052731), cho thấy mô hình mắc ít lỗi gán sai nhãn nhất.
  + Đạt Jaccard Accuracy cao nhất (0.5935), phản ánh mức độ trùng khớp tốt giữa tập nhãn dự đoán và tập nhãn thực tế.
  + Đồng thời, có micro-F1 đứng thứ hai (0.62), chỉ thấp hơn mô hình tốt nhất khoảng 2.8%, điều này chứng minh tính hiệu quả và thực tiễn của mô hình nhẹ DistilBERT khi tối ưu hóa chi phí tính toán.
- Trong khi đó, mô hình BERT-base-cased huấn luyện trên dữ liệu gốc (raw + trimmed) đạt micro-F1 thấp nhất (0.619079) và Jaccard accuracy thấp nhất (0.56525), cho thấy rằng việc không tiền xử lý có thể khiến mô hình chịu ảnh hưởng bởi nhiễu văn bản.

4.7.3. Kết quả mô hình MultiOutput Logistic Regression

Dataset	**Model**	**Multi-label Classification**		
		micro-F1	**Hamming Loss**	**Jaccard Accuracy**
Preprocessed	Logistic Regression	0.46	0.066	0.3608

Sau khi thực hiện đánh giá trên tập kiểm thử dùng mô hình MultiOuput Logistic Regression, nhóm đề tài nhận thầy rằng mặc dù đã áp dụng các kỹ thuật tiền xử lý mạnh nhưng kết quả vẫn thấp hơn đáng kể so với các mô hình họ BERT.

- Chỉ số micro-F1 = 0.46 → Phản ánh độ chính xác tổng thể thấp hơn so với các mô hình họ BERT.

- Chỉ số Hamming Loss = 0.066 → Cao nhất trong tất cả các mô hình, cho thấy rằng mô hình thường xuyên gán nhãn sai nhiều hơn so với các mô hình khác.
- Chỉ số Jaccard accuracy = 0.3608 → Cho thấy sự khác biệt lớn giữa tập nhãn dự đoán và tập nhãn thực.

Dễ dàng thấy rằng, kết quả trên micro-F1, Hamming Loss và Jaccard accuracy của các mô hình họ BERT vượt trội hơn so với mô hình máy học cơ bản là Logistic Regression. Điều này là do các mô hình họ BERT đã được tiền huấn luyện trên một dữ liệu lớn, có khả năng học ngữ cảnh sâu và hiểu được mối quan hệ ngữ nghĩa phức tạp của văn bản, đặc biệt là đối với các văn bản mô tả phim thường có nhiều biểu hiện ngữ nghĩa dài, giàu cấu trúc, bao gồm tên riêng, thành ngữ, hoặc diễn đạt gián tiếp. Tuy nhiên, các mô hình họ BERT thường khá phức tạp, nên nó sẽ đòi hỏi nhiều tài nguyên hơn và thời gian dự đoán sẽ lâu hơn so với các mô hình học máy cơ bản.

Mô hình Logistic Regression có thể hữu ích như một baseline tham chiếu, nhưng không phù hợp để khai thác các đặc trưng ngữ nghĩa sâu trong bài toán phân loại đa nhãn với dữ liệu văn bản tự nhiên. Trong khi đó, các mô hình họ BERT — đặc biệt là BERT-base-uncased và DistilBERT — thể hiện hiệu suất vượt trội rõ rệt, ngay cả trên dữ liệu đã được cắt gọn hoặc tiền xử lý mạnh.

4.8. Phân tích lỗi (Error Analysis)

Mô hình	Nội dung ngữ liệu	Nhãn thực	Nhãn dự đoán	Phân tích
BERT-base-uncased	Paul Atreides unites with the Fremen while on a warpath of revenge against the conspirators who destroyed his family. Facing a choice between the love of his life and the fate of the universe, he endeavors to prevent a terrible future.	Action Adventure Drama Science Fiction	Action Adventure Science Fiction	Thiếu nhãn Drama. Văn bản mô tả một cuộc trả thù quy mô lớn, liên quan đến định mệnh và tình yêu, điều này thể hiện yếu tố tâm lý/phát triển nhân vật (Drama). Mô hình có thể bỏ sót yếu tố cảm xúc nội tâm vì tập huấn luyện ưu tiên ngữ cảnh "chiến tranh, không gian". Các nhãn Action, Adventure, Science Fiction đều đúng, vì văn bản chứa các yếu tố chiến đấu, phiêu lưu, và bối cảnh vũ trụ.

	While scavenging the deep ends of a derelict space station, a group of young space colonists come face to face with the most terrifying life form in the universe.	Science Fiction Thriller Horror	Science Fiction Horror	Dự đoán được nhãn Science Fiction nhưng thiếu nhãn Thriller. Dù có cảm giác hồi hộp/đe dọa, mô hình không gán nhãn "Thriller" → cho thấy mô hình có thể khó phân biệt ranh giới giữa Horror và Thriller khi nội dung thiên về quái vật hoặc môi trường cô lập.
	Mufasa, a cub lost and alone, meets a sympathetic lion named Taka, the heir to a royal bloodline. The chance meeting sets in motion an expansive journey of a group of misfits searching for their destiny.	Adventure Music Drama Family Animation	Family Adventure Animation	Thiếu nhãn Music. Có thể không có từ khóa rõ ràng về âm nhạc trong văn bản – trừ khi phần nội dung gốc đề cập cụ thể đến ca hát, nhạc nền. Thiếu nhãn Drama. Có thể vì mô hình chưa nhận diện diễn biến cảm xúc nội tâm/bi kịch qua cụm như “searching for their destiny”.
	A sudden attack by Wulf, a clever and traitorous lord of Rohan seeking vengeance for the death of his father, forces Helm Hammerhand, the King of Rohan, and his people to make a daring last stand in the ancient stronghold of the Hornburg.	Adventure, Action, Animation	Adventure, Action, Fantasy	Dự đoán được nhãn Adventure và Action. Thiếu nhãn Animation vì miêu tả không đề cập trực tiếp đến yếu tố hình ảnh hoạt hình. Dự đoán thêm nhãn Fantasy hợp lý vì mô tả bối cảnh vương quốc, vua chúa, chiến tranh cổ xưa rất gần với thể loại Fantasy.
	After being betrayed by his master, a disillusioned assassin takes one final shot	Action, Thriller	Crime, Action, Thriller	Nhãn Crime được mô hình thêm vào hợp lý vì có yếu tố tội phạm, phản bội, ám sát.

	with the intention of using the money to help a musician he accidentally blinded regain vision.			
distilBERT	Elio, a space fanatic with an active imagination, finds himself on a cosmic misadventure where he must form new bonds with alien lifeforms, navigate a crisis of intergalactic proportions and somehow discover who he is truly meant to be.	Science Fiction Animation Drama Fantasy Family	Comedy, Animation, Science Fiction	Dự đoán được Animation và Science Fiction, đồng thời gán nhầm Comedy. Việc bỏ sót các nhãn như Family hay Fantasy cho thấy mô hình thiếu khả năng nhận diện các yếu tố nhẹ nhàng, cảm xúc và phù hợp với trẻ em, do không xuất hiện rõ qua từ khóa trong mô tả.
	Two priests, one in crisis with his faith and the other confronting a turbulent past, must overcome their differences to perform a risky exorcism.	Drama, Horror, Thriller	Horror, Document ary, Drama	Dự đoán đúng nhãn Horror. Từ khóa như exorcism, crisis, turbulent past gợi ra yếu tố tâm lý căng thẳng, phù hợp với Thriller, tuy nhiên mô hình phát hiện nhãn này. Documentary có thể do bối cảnh nghiêm túc và tín ngưỡng khiến mô hình nhầm với phim tài liệu về tôn giáo.
	An assassin trained in the traditions of the Ruska Roma organization sets out to seek revenge after her father's death.	Action, Thriller	Comedy, Action, Drama	Nhãn đúng là Action, Thriller, nhưng mô hình chỉ nhận ra Action, đồng thời gán thêm Comedy và Drama. Comedy là nhãn sai hoàn toàn, không có dấu hiệu hài hước trong mô tả. Mô hình cũng bỏ sót Thriller, thể hiện hạn chế trong việc nhận diện nhịp độ cao và yếu tố căng thẳng

				thường thấy trong phim hành động - báo thù.
	In April 1986, the city of Chernobyl in the Soviet Union suffers one of the worst nuclear disasters in the history of mankind. Consequently, many heroes put their lives on the line in the following days, weeks and months.	Thriller, Drama, History	Document ary, Drama, Comedy	Dự đoán đúng nhãn Drama. Thiếu nhãn Thriller và History. Documentary là hợp lý nếu xét theo nội dung overview là về thảm họa hạt nhân Chernobyl. Mô hình gán nhãn Comedy là lỗi nặng vì nội dung không liên quan đến hài.
	The British military recruits a small group of highly skilled soldiers to strike against German forces behind enemy lines during World War II.	Action, Comedy, War	War, History, Drama	Mô hình nghiêng về thể loại lịch sử nghiêm túc, không nhận ra yếu tố hành động và hài nếu phim mang phong cách phiêu lưu giải trí. Mô hình dự đoán đúng nhãn War, bỏ sót nhãn Action và Comedy nhưng thêm được nhãn History và Drama
BERT-base -cased	When all but one child from the same class mysteriously vanish on the same night at exactly the same time, a community is left questioning who or what is behind their disappearance.	Horror, Mystery	Horror, Thriller	Mô hình đoán đúng không khí rùng rợn Horror, nhưng thay vì nhận ra yếu tố bí ẩn Mystery, lại chọn Thriller. Đây là lỗi phổ biến do hai nhãn này thường bị nhầm nếu mô hình tập trung vào nhịp độ hoặc căng thẳng.
	A chronicle of New York's drag scene in the 1980s, focusing on balls, voguing and the ambitions and dreams	Document ary	TV Movie, Comedy	Mô hình sai hoàn toàn: thiếu nhãn Documentary, thay vào đó là TV Movie và Comedy. Có thể do từ khóa như “balls”, “ambitions”, “vitality” khiến

	of those who gave the era its warmth and vitality.			mô hình hiểu sai thành phim giải trí thay vì tài liệu, cho thấy sự hạn chế trong hiểu ngữ cảnh văn hóa xã hội.
	A seventeen-year-old aristocrat falls in love with a kind but poor artist aboard the luxurious, ill-fated R.M.S. Titanic.	Drama, Romance	Romance, Music, Adventure	Mô hình giữ đúng Romance, nhưng thay Drama bằng Music, Adventure, cho thấy sự nhiễu khi gặp các từ như "artist", "luxurious", "ill-fated". Music không có cơ sở rõ ràng; Adventure cũng không phù hợp vì đây là bi kịch tình cảm.
	When oil is discovered in 1920s Oklahoma under Osage Nation land, the Osage people are murdered one by one - until the FBI steps in to unravel the mystery.	Crime, Drama, History, Western	Thriller, Action, Crime	Mô hình chỉ giữ được Crime, nhưng lại bỏ mất History và Western – hai yếu tố đặc trưng của bối cảnh. Action là nhãn sai, vì phim nghiêng về điều tra hơn là hành động. Có thêm nhãn Crime là hợp lý vì có các từ như “murder”, “unravel”
	A promising young drummer enrolls at a cut-throat music conservatory where his dreams of greatness are mentored by an instructor who will stop at nothing to realize a student's potential.	Music, Drama	Music, Comedy, Drama	Music và Drama đều đúng, nhưng thêm Comedy là sai, có thể do mô hình hiểu nhầm các từ liên quan đến “giấc mơ”, “năng lượng” hoặc không phân biệt được tông phim căng thẳng từ ngữ cảnh mô tả.

Trong quá trình thử nghiệm mô hình, ba mô hình ngôn ngữ gồm distilBERT, bert-base-cased và bert-base-uncased trong bài toán phân loại đa nhãn thể loại phim từ mô tả (overview), có thể nhận thấy một số xu hướng lỗi của mô hình. Cả ba mô hình đều thể hiện khả năng nhận diện tốt các thể loại phổ biến có từ khóa rõ ràng như Action, Adventure, Science Fiction hoặc Drama, nhưng lại gặp khó khăn đáng kể với

các thể loại mang tính trừu tượng, tâm lý hoặc đòi hỏi tri thức như Family, Romance, History hay Documentary. Một lỗi phổ biến là việc bỏ sót các nhãn liên quan đến cảm xúc nội tâm, đặc biệt là Drama, do mô hình thường không nhận diện được chiều sâu tâm lý nếu không có từ khóa trực tiếp trong văn bản. Ngoài ra, các mô hình cũng hay nhầm lẫn giữa Mystery và Thriller, phản ánh sự thiếu tinh tế trong việc phân biệt giữa sắc thái hồi hộp và khám phá bí ẩn.

Bên cạnh đó, các mô hình cũng cho thấy xu hướng gán nhãn sai cho một số thể loại mang tính ngữ dụng cao. Điển hình là việc Comedy, Adventure hoặc TV Movie bị gán nhầm cho những văn bản có giọng điệu tích cực, cấu trúc hành trình hoặc đề cập đến "misadventures" – ngay cả khi nội dung thực chất mang tính bi kịch hoặc tâm lý nặng. Đáng chú ý, các mô hình thường xuyên bỏ sót các nhãn phản ánh bối cảnh lịch sử hoặc xã hội như History, Western hay Documentary, ngay cả khi overview có nhắc đến thời gian, địa danh hoặc sự kiện thực tế. Điều này cho thấy mô hình còn thiếu khả năng liên kết ngữ cảnh ngôn ngữ với tri thức thế giới (world knowledge), một yếu tố đặc biệt quan trọng khi xử lý văn bản mang tính văn hóa hoặc lịch sử.

So sánh giữa các mô hình, bert-base-cased cho thấy khả năng nhận diện tốt hơn đối với các thể loại nghiêm túc như Horror và Romance, trong khi bert-base-uncased lại thể hiện ưu thế trong việc xử lý các văn bản mang tính hành động hoặc giả tưởng. distilBERT, mặc dù nhẹ hơn về mặt tính toán, lại có xu hướng bỏ sót các thể loại tâm lý và gán sai các nhãn hài hoặc phiêu lưu. Tựu trung, cả ba mô hình đều phụ thuộc nhiều vào biểu hiện ngôn ngữ bề mặt và thiếu cơ chế hiểu sâu về cấu trúc cốt truyện, tiến trình cảm xúc và tri thức nền. Để nâng cao hiệu quả phân loại, cần cân nhắc tích hợp thêm các đặc trưng ngữ nghĩa chuyên sâu (semantic roles, discourse structure), mô hình hóa cảm xúc, hoặc huấn luyện với dữ liệu được chú thích đa tầng, đồng thời áp dụng các kỹ thuật tổ hợp mô hình (ensemble learning).

Chương 5. KẾT LUẬN

5.1. Các điểm nổi bật rút ra từ quá trình thực hiện đề tài

Qua quá trình nghiên cứu và thực nghiệm, đề tài đã đạt được những kết quả và rút ra được các điểm nổi bật sau:

- **Tầm quan trọng của tiền xử lý dữ liệu**: Quá trình tiền xử lý văn bản một cách kỹ lưỡng (loại bỏ nhiễu, chuẩn hóa, lemmatization) và xử lý mất cân bằng dữ liệu (sử dụng phương pháp cắt mẫu) đã giúp mô hình cải thiện hiệu suất

- **Hiệu quả của kiến trúc Transformer**: Các mô hình dựa trên kiến trúc Transformer (họ BERT) cho thấy hiệu suất vượt trội rõ rệt so với phương pháp học máy truyền thống (Logistic Regression kết hợp TF-IDF). Mô hình BERT-base-uncased trên dữ liệu đã tiền xử lý và cắt mẫu đạt điểm micro-F1 cao nhất (0.648), khẳng định khả năng học ngữ cảnh sâu của các mô hình ngôn ngữ lớn.

- **Sự cân bằng giữa hiệu suất và tốc độ xử lý**: Mô hình DistilBERT nổi bật như một lựa chọn tối ưu, đạt được hiệu suất rất cao (Hamming Loss thấp nhất 0.052 và Jaccard Accuracy cao nhất 0.593) trong khi có kích thước nhỏ gọn và tốc độ dự đoán nhanh hơn. Điều này chứng tỏ tính thực tiễn của các mô hình này có thể được sử dụng cho các ứng dụng thực tế.

- **Phân tích lỗi sâu sắc**: Việc phân tích lỗi chi tiết đã chỉ ra những điểm mạnh và điểm yếu cố hữu của các mô hình. Các mô hình nhận diện tốt các thể loại có từ khóa rõ ràng (ví dụ: "Action" hay "Science Fiction") nhưng gặp khó khăn với các thể loại trừu tượng, đòi hỏi suy luận về cảm xúc (ví dụ: "Drama", "History", "Documentary").

5.2. Đóng góp của đề tài

Từ kết quả dự đoán, nhóm có thể cho thấy được, mô hình của nhóm có thể áp dụng trong các hệ thống phim. Thay vì trước đó, phân loại các thể loại phim thủ công, có thể áp dụng hệ thống vào để tự động hóa việc đánh nhãn. Đề tài cũng cho thấy điểm yếu của mô hình học máy truyền thống khi so sánh với kiến trúc Transformer trong bài toán đa đầu ra.

5.3. Hạn chế và hướng phát triển

Bên cạnh những mặt tốt, đề tài của nhóm vẫn tồn tại những điểm yếu nhất định:

- Mô hình vẫn chưa thực sự đạt độ chính xác cao khi dự đoán những mô tả phim có nhãn mang tính trừu tượng cao, có bối cảnh, ngữ nghĩa và diễn tiến cảm xúc phức tạp,

- Nhóm đề tài chưa tận dụng hết được sức mạnh của mô hình họ BERT khi mới chỉ áp dụng theo hướng cơ bản, chưa khai thác hết các kỹ thuật xử lý ngôn ngữ.

- Mặc dù sở hữu tập dữ liệu lớn, nhóm mới chỉ sử dụng phần mô tả phim (overview) để huấn luyện mô hình, trong khi nhiều thông tin bổ trợ khác chưa được khai thác. Bên cạnh đó, sự mất cân bằng giữa các lớp nhãn cũng ảnh hưởng tiêu cực đến hiệu quả huấn luyện, khiến mô hình dễ thiên lệch về các thể loại phổ biến và bỏ sót các thể loại ít xuất hiện.

Từ những mặt hạn chế trên, nhóm đã có định hướng phát triển đề tài cho sau này:

- Từ bộ dữ liệu gốc, nhóm dự kiến sẽ khai thác thêm một số đặc trưng khác như tên phim (title), đánh giá phim (rating), số lượt bình chọn (vote count) hoặc độ dài phim (runtime) nhằm bổ sung ngữ cảnh và tăng độ phong phú cho đầu vào của mô hình. Những đặc trưng này có thể góp phần hỗ trợ mô hình phân biệt tốt hơn giữa các thể loại có nội dung gần nhau, đồng thời giúp cải thiện khả năng dự đoán ở các trường hợp thiếu thông tin rõ ràng trong phần mô tả (overview).

- Sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên nâng cao như text augumentation (tăng cường dữ liệu văn bản) nhằm cải thiện khả năng tổng quát hóa của mô hình. Cụ thể, các phương pháp như thay thế từ đồng nghĩa, hoán vị câu, sinh paraphrase bằng mô hình ngôn ngữ, hoặc dịch ngược (back-translation) sẽ được sử dụng để tạo thêm các biến thể hợp lệ của mô tả phim (overview) .Từ đó giúp cân bằng, làm tăng tính đa dạng của dữ liệu và giúp mô hình học được nhiều dạng biểu đạt nội dung khác nhau cho cùng một thể loại.

- Ngoài ra, nhóm cũng xem xét giảm số lượng các lớp gây nhập nhằng dữ liệu (ambiguous class) bằng cách gộp các thể loại có ý nghĩa gần nhau hoặc xuất hiện với tần suất thấp vào các nhóm chung. Cách tiếp cận này giúp giảm độ phức tạp của bài toán phân loại, hạn chế hiện tượng chồng lấn giữa các nhãn (label overlap), đồng thời cải thiện độ ổn định của mô hình trong việc nhận diện các thể loại khó phân biệt.

- Bên cạnh các cải tiến về dữ liệu và tiền xử lý, nhóm cũng định hướng huấn luyện thêm trên các kiến trúc mô hình khác ngoài họ BERT, chẳng hạn như RoBERTa, DeBERTa, hoặc các mô hình tối ưu hóa cho đa nhãn như X-Transformer. Việc so sánh hiệu suất giữa các mô hình này sẽ giúp nhóm đánh giá mức độ phù hợp giữa kiến trúc và đặc thù của bài toán phân loại thể loại phim.

5.4. Tổng kết.

Đề tài đã tập trung xây dựng mô hình phân loại đa nhãn thể loại phim dựa trên mô tả nội dung (overview) bằng cách ứng dụng các mô hình ngôn ngữ họ BERT như bert-base-cased, bert-base-uncased và distilBERT. Kết quả thực nghiệm cho thấy mô hình có khả năng nhận diện tốt các thể loại phổ biến và rõ ràng về mặt ngôn ngữ, tuy nhiên vẫn gặp khó khăn với các thể loại trừu tượng, giàu cảm xúc hoặc mang tính văn hóa – lịch sử. Đề tài cũng đã đánh giá hiệu suất thông qua các chỉ số như micro-F1, Hamming Loss, và Jaccard accuracy đồng thời chỉ ra những sai lệch phổ biến trong dự đoán và nguyên nhân tiềm ẩn.

Từ những kết quả đạt được, nhóm đã xác định một số định hướng phát triển như: mở rộng tập đặc trưng đầu vào (thêm title, rating, vote count...), áp dụng các kỹ thuật tăng cường dữ liệu văn bản (text augmentation), xử lý mất cân bằng nhãn, giảm số lượng lớp gây nhập nhằng, cũng như thử nghiệm trên các kiến trúc mô hình khác. Những định hướng này không chỉ góp phần nâng cao chất lượng mô hình mà còn đặt nền tảng cho các nghiên cứu mở rộng sau này, như xây dựng hệ thống gợi ý thể loại phim tự động hoặc ứng dụng trong khai phá sở thích người dùng.

Chương 6. CÁC TÀI LIỆU THAM KHẢO

[1] wykonos, *"Original dataset wykonos/movies"*, HuggingFace, 2025. [Online]. Available: https://huggingface.co/datasets/wykonos/movies

[2] Wikipedia, *"Thể loại phim – Wikipedia tiếng Việt"*, 2025. [Online]. Available: https://vi.wikipedia.org/wiki/Thể_loại_phim

[3] PyTorch, *"BCEWithLogitsLoss — PyTorch 2.7 documentation"*, 2025. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

[4] H. Đ. Quân, *"Giải thuật Jaccard"*, 2025. [Online]. Available: https://viblo.asia/p/giai-thuat-jaccard-djeZ1P9GKWz

[5] S. Lee, *"Hamming Loss Explained: Key Insights for Multi-label Learning"*, 2025. [Online]. Available: https://www.numberanalytics.com/blog/hamming-loss-explained-key-insights

[6] GeeksforGeeks, *"BERT Model - NLP"*, 2025. [Online]. Available: https://www.geeksforgeeks.org/bert-model/

[7] Trí tuệ nhân tạo, *"Bài 6: Logistic Regression (Hồi quy Logistic)"*, 2025. [Online]. Available: https://trituenhantao.io/machine-learning-co-ban/bai-6-logistic-regression-hoi-quy-logistic/

[8] GeeksforGeeks, *"Ensemble Learning"*, 2025. [Online]. Available: https://www.geeksforgeeks.org/ensemble-learning/