

Movie Genre Classification From Overview

About us



Minh Chí



Minh Chiến



Phương Thảo

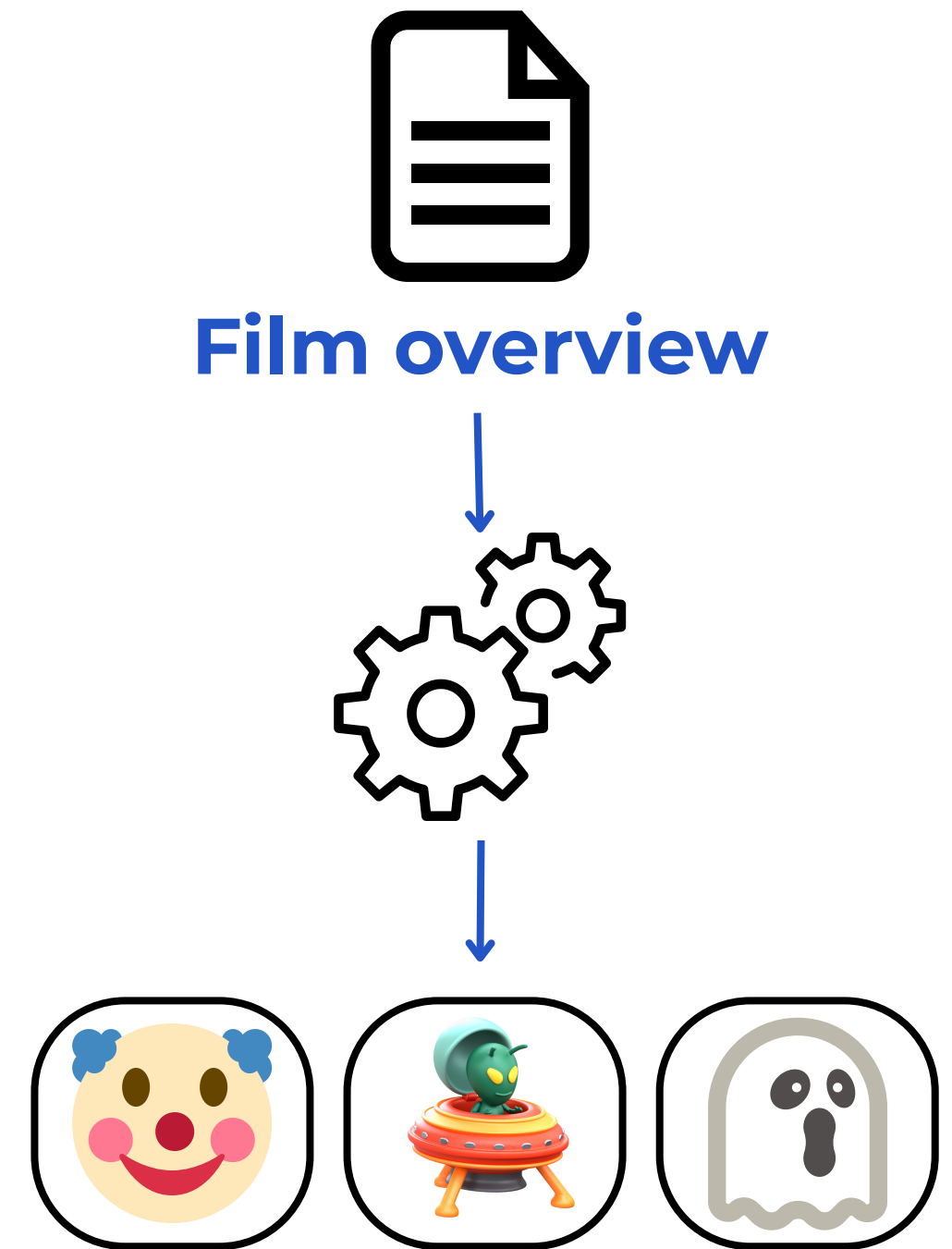
We are PhongBART

Giới thiệu bài toán

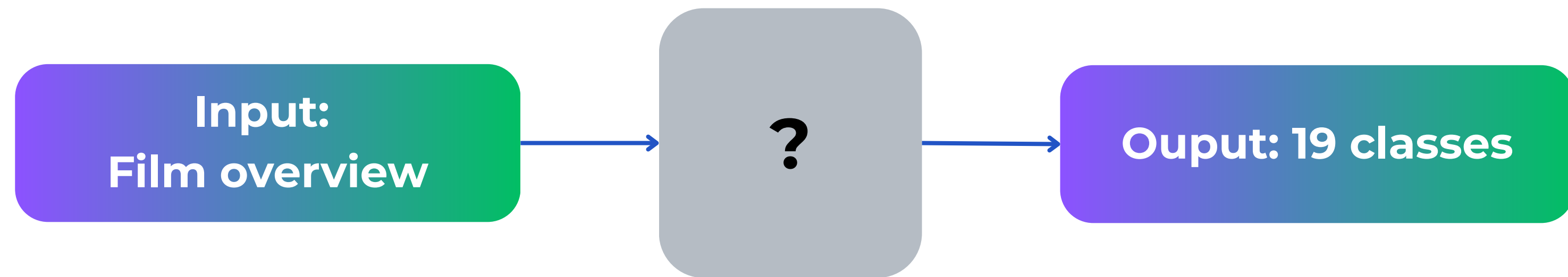
Thực trạng:

- Việc phân loại các phim trở nên **thiết yếu**
- Phân loại thủ công có thể **mất nhiều thời gian**
- Giúp các **hệ thống** đề xuất phim phù hợp

→ Làm thế nào để **tự động hóa** việc **phân loại phim** dựa trên một **miêu tả tổng quan của một bộ phim**?



Giới thiệu bài toán



Horror

War

Adventure

Drama

Western

Romance

Comedy

Music

Mystery

History

TV Movie

Crime

Animation

Sci-fi

Fantasy

Family

Documentary

Thriller

Action

Ví dụ dữ liệu

Input:

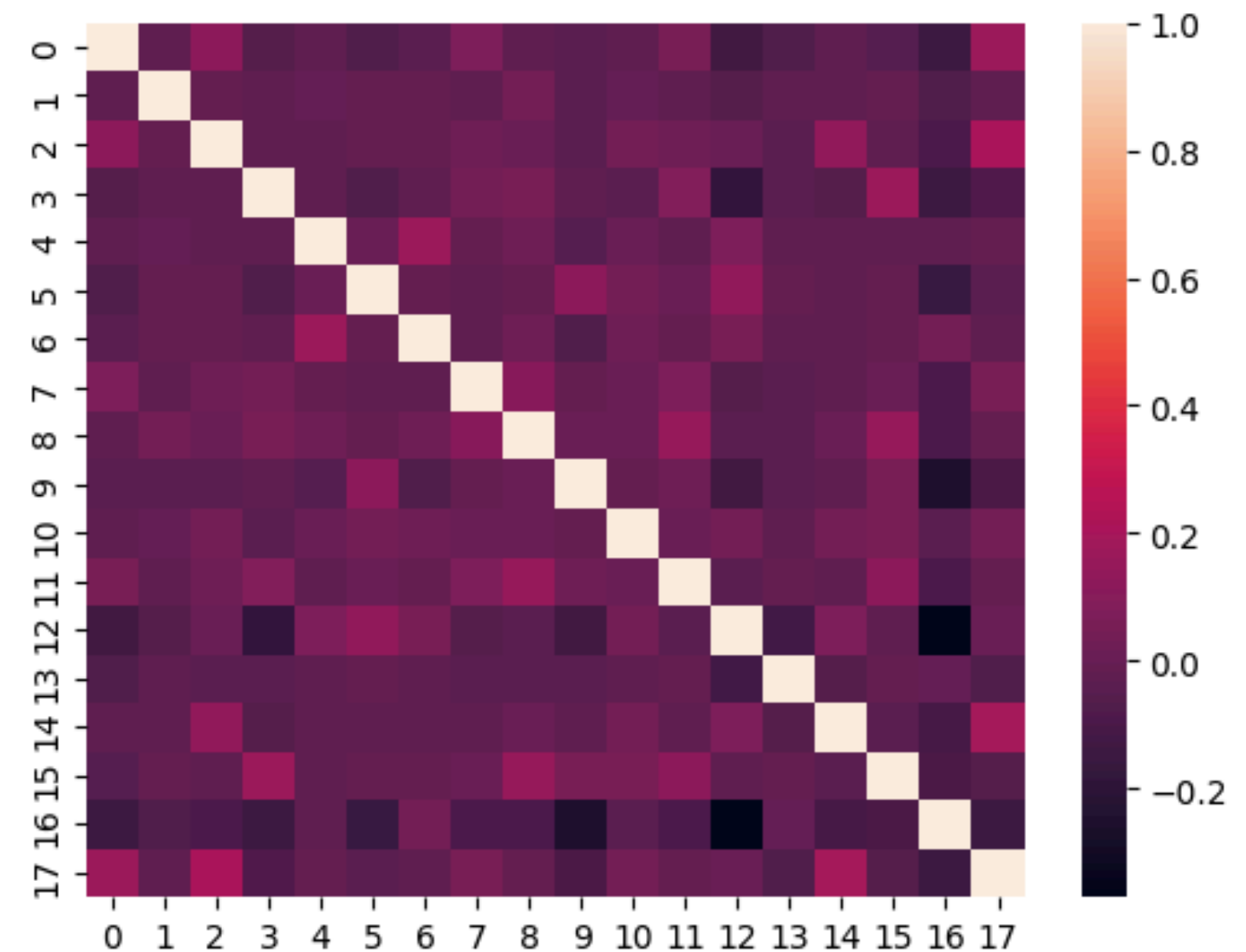
Alex an assassin-for-hire finds that he's **become a target** after he refuses to complete a job for a **dangerous criminal organization**. With the crime syndicate and **FBI** in hot pursuit Alex has the skills to stay ahead except for one thing: he is struggling with severe memory loss affecting his every move. Alex must question his every action and whom he can ultimately trust.

Output:

Crime
Action
Thriller

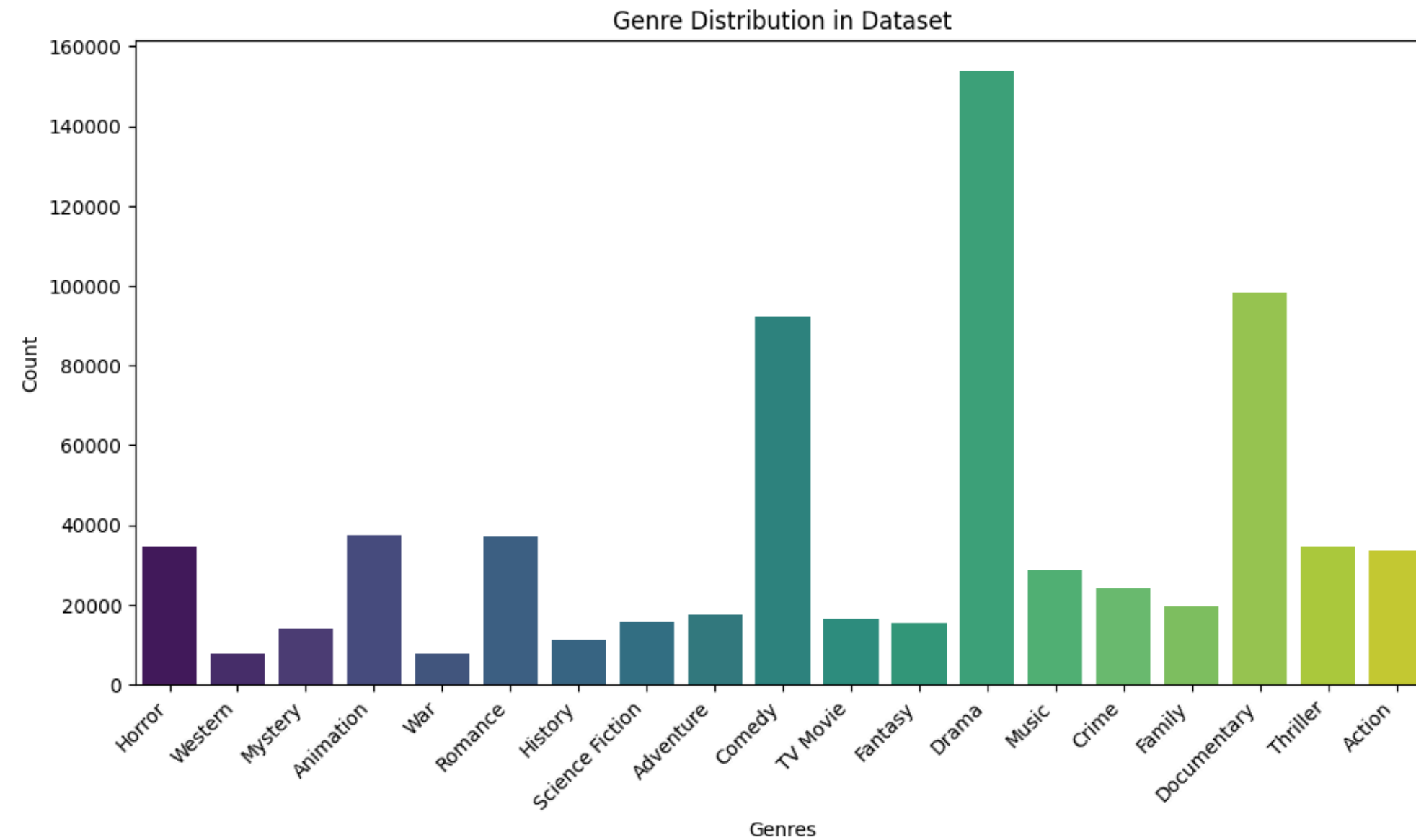
Tổng quan ngữ liệu

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



Mã trận tương quan giữa các lớp

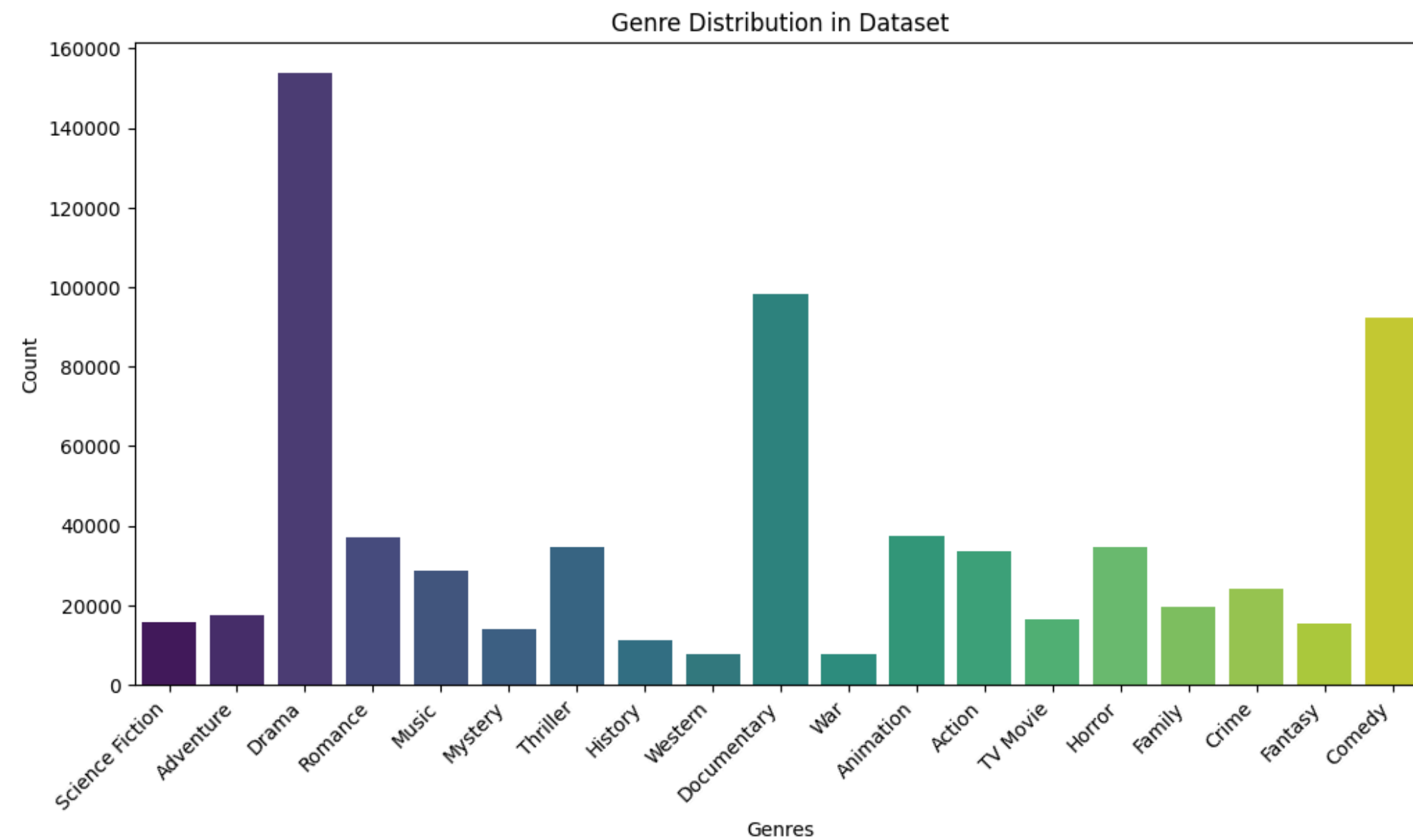
EDA



Number of samples:
722,796

Raw data

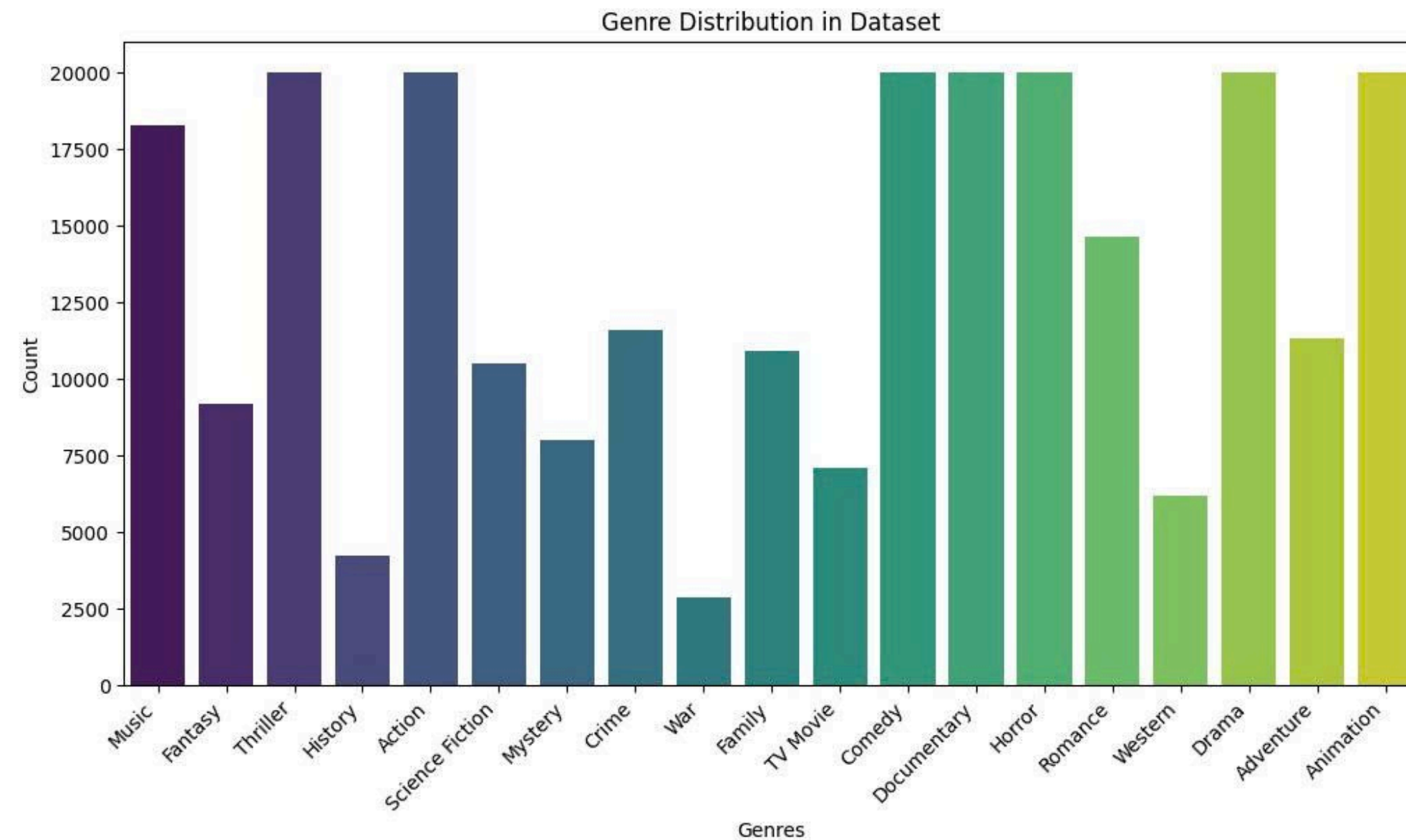
EDA



Number of samples:
722,796 → **435,706**

Data after preprocessing

EDA



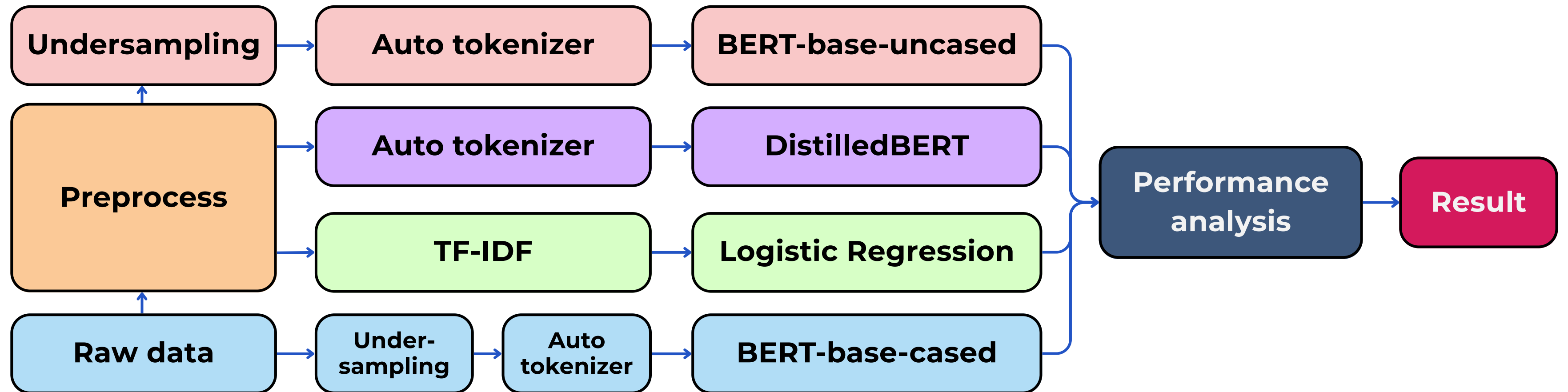
Number of samples:
722,796 → **435,706** → **157,160**

Data after preprocessing and undersampling

Phân tích dữ liệu

Dữ liệu	Nhãn	Phân tích
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.
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 quá rõ về đặc trưng của nhãn War. Chưa có yếu tố nào để xác định đây là nhãn Drama.
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.

Sơ đồ hóa các bước

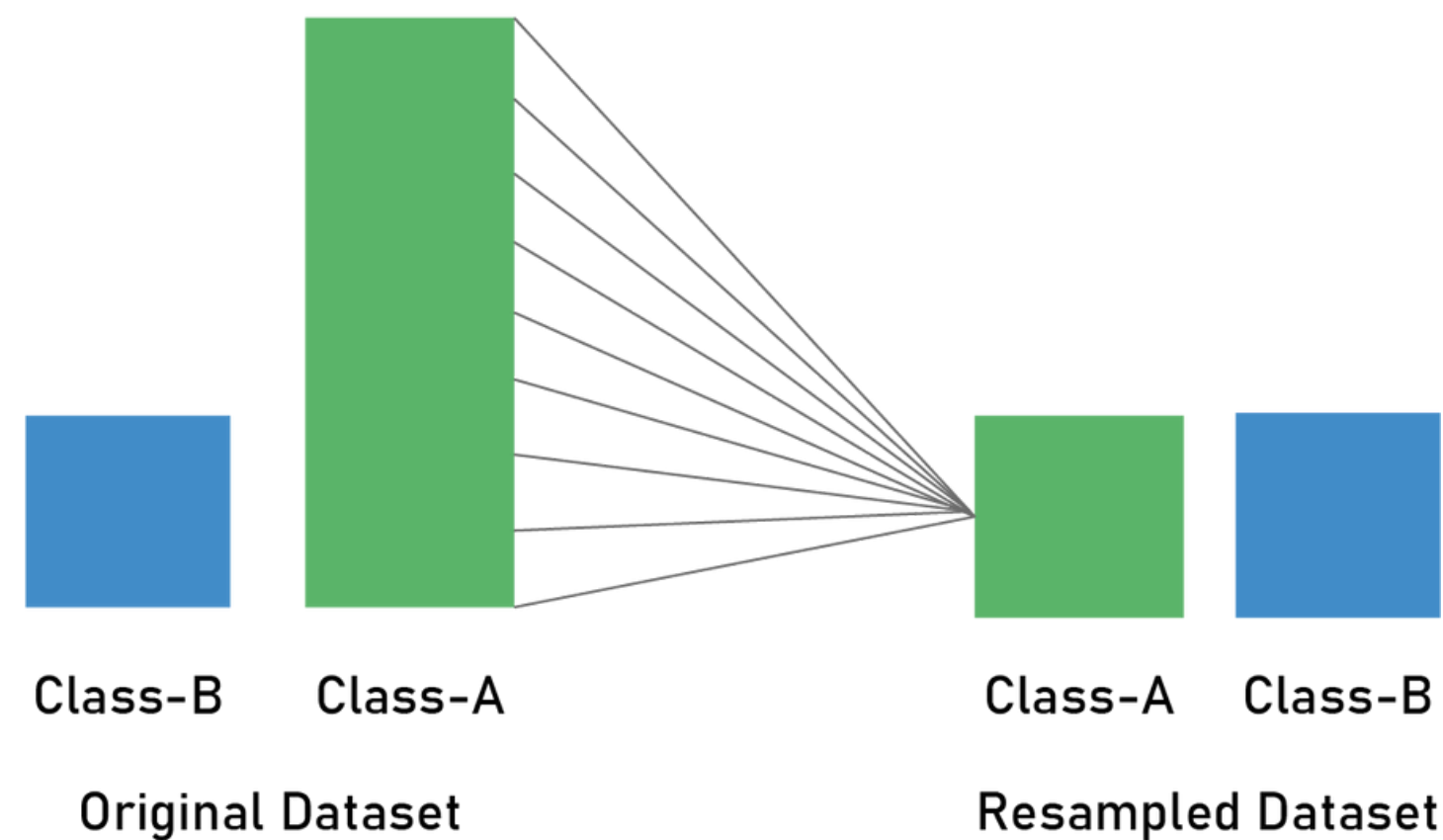


Các bước tiền xử lý dữ liệu

- Áp dụng **One-hot encoding** để biểu diễn label **đa lớp**.
 - **Loại bỏ** các kí tự **HTML**.
 - **Loại bỏ** các **liên kết web**.
 - **Thay đổi** các **từ viết gọn, viết tắt** thành **từ hoàn chỉnh**.
 - Chuyển tất cả về **ký tự in thường**.
 - **Loại bỏ chữ số** và **dấu câu**.
 - **Loại bỏ** các **emoji** và **emoticon**.
 - **Tokenize**.
 - Loại bỏ **stopword**.
 - Áp dụng **lemmatize**.
-

Cắt giảm số lượng mẫu dữ liệu

Under Sampling



- Áp dụng phương pháp **Undersampling** để cân bằng lại số lượng mẫu giữa các lớp.
- Giúp tăng độ chính xác trên các lớp với số lượng sample hạn chế.

Huấn luyện mô hình

Data processing:

- Tokenizer: AutoTokenizer
- Dataset split: 70:20:10

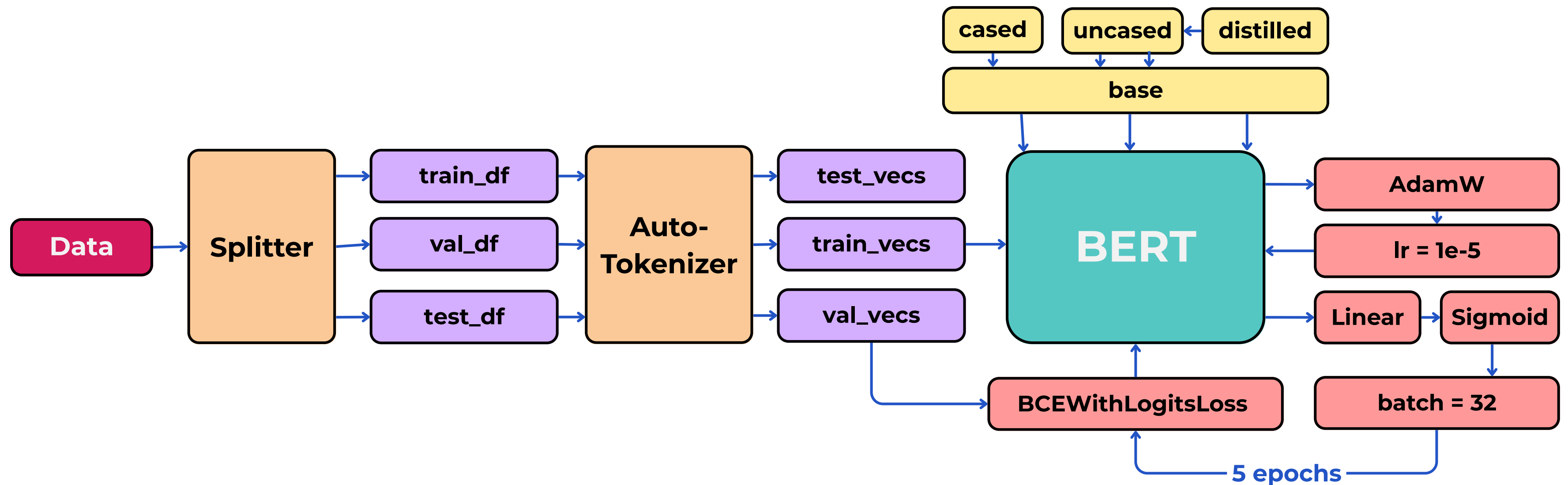
Hyperparameters:

- Optimizer: AdamW
- Learning rate: $1e-5$
- Epochs: 5
- Batch Size: 32

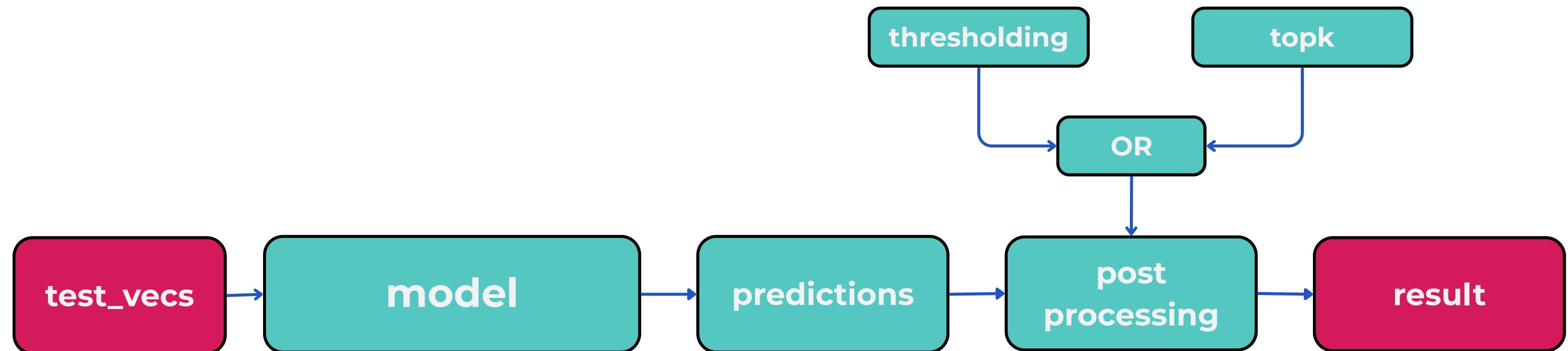
Model architecture:

- Base model:
 - BERT-base-uncased
 - BERT-base-cased
 - DistilledBERT
 - Final layer: Linear (768 → 19)
 - Activation: Sigmoid
 - Loss: BCEWithLogitsLoss
-

Huấn luyện mô hình



Huấn luyện mô hình



Đánh giá mô hình


- Sử dụng các metric gồm: micro-F1, Jaccard accuracy và Hamming Loss

Dataset	Model	Multi-label Classification		
		micro-F1	Hamming Loss	Jaccard Accuracy
Raw + trimmed	Bert-base-cased	0.619079	0.053693	0.56525
Preprocessed	Logistic Regression	0.46	0.068	0.3587
	DistilBERT	0.62	0.052731	0.5935
Preprocessed + trimmed	Bert-base-uncased	0.648779	0.063554	0.589834

Thử nghiệm mô hình

Dữ liệu	Thực	Dự đoán	Phân tích
What starts out as girls weekend away in the Mojave desert becomes a tale of horror, death and alien invasion.	Science Fiction, Horror	Science Fiction, Horror	Các từ ngữ “horror”, “death” thể hiện rõ nhãn Horror. Cụm “alien invasion” phù hợp với nhãn Science Fiction.
In a series of escalating encounters, former security guard David Dunn uses his supernatural abilities to track Kevin Wendell Crumb, a disturbed man who has twenty-four personalities. Meanwhile, the shadowy presence of Elijah Price emerges as an orchestrator who holds secrets critical to both men.	Drama, Thriller, Science Fiction	Horror, Thriller	Các cụm từ “twenty-four personalities”, “hold secrets critical to..” mang lại cảm giác tò mò bí ẩn hợp với nhãn Thriller. Mô tả có cảm giác gay cấn cùng các chi tiết “supernatural abilities” phù hợp với nhãn Drama và Science Fiction, tuy nhiên bị nhầm lẫn với nhãn Horror.
In Jeju, a spirited girl and a steadfast boy's island story blossoms into a lifelong tale of setbacks and triumphs, proving love endures across time.	Drama Romance History	Drama, Romance	Nội dung nói về tình yêu bền bỉ sau bao lần bị cản trở tách biệt và gặp lại, phù hợp với nhãn Drama và Romance. Tuy nhiên trong overview lại không có yếu tố lịch sử History. Mô hình dự đoán được 2/3 nhãn thực tế.

Streamlit

 **Movie Genre Prediction from Overview**

Enter the movie overview:

Doraemon and friends go on an adventure to meet new buddies, connect to people with music, and save the world from a crisis.

Choose a base model:

bert-base-uncased

Predict Genres

Prediction Complete

Top Predicted Genres: ↗

Animation: 0.98

Family: 0.89

Thực nghiệm mô hình trên streamlit

Kết quả dự đoán

Mở rộng đề tài

- Chọn thêm **các features khác** để đánh giá
 - Sử dụng các phương pháp **text augmentation**
 - **Giảm** số lượng **ambiguous class**
 - Huấn luyện thêm trên **các kiến trúc khác**
 - Thực hiện **ensemble** trên các mô hình có kết quả chính xác cao
-

Tài liệu tham khảo

1. DistiledBERT Documents,
https://huggingface.co/docs/transformers/en/model_doc/distilbert
 2. Multi-label text classification using BERT,
<https://github.com/dtolk/multilabel-BERT>
 3. Movie dataset,
<https://huggingface.co/datasets/wykonos/movies>
 4. Streamlit documentation,
<https://docs.streamlit.io>
-

Thank You

We are ready to assist you



University of Information Technology, VNUHCM
