

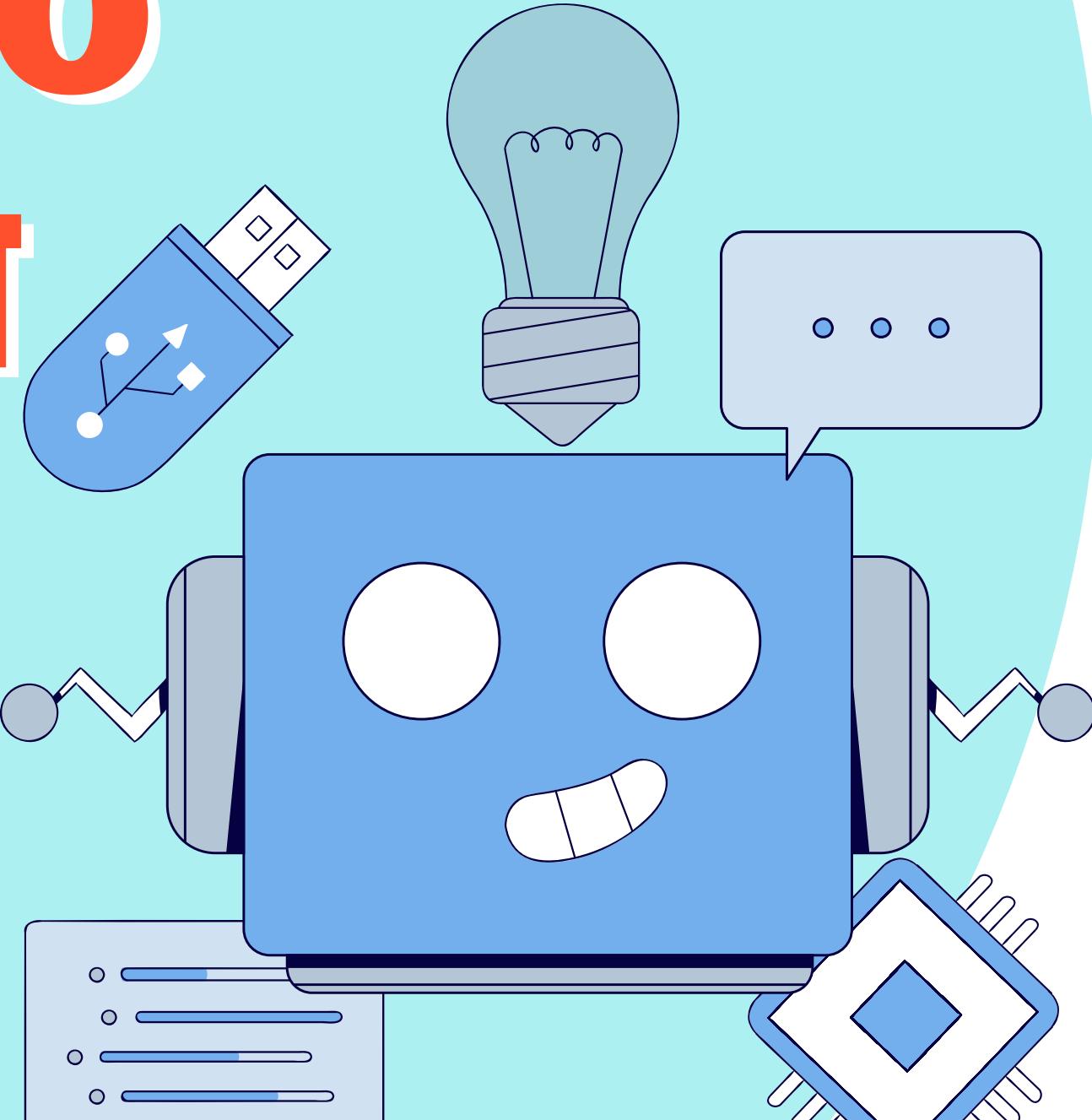
DS105



Khoa Khoa học
và Kỹ thuật Thông tin

NHÓM 17

PHÂN TÍCH YẾU TỐ ẢNH HƯỞNG ĐẾN DOANH THU PHIM



THÀNH VIÊN NHÓM



22521207
**Trương Nhật
Quang**



22521548
Đỗ Tuấn Trực



22521277
**Nguyễn Đức
Tài**



22521294
**Trương Lưu
Song Tâm**

MỤC LỤC:



GIỚI THIỆU



MÔ TẢ BỘ DỮ LIỆU



PHƯƠNG PHÁP PHÂN TÍCH



PHÂN TÍCH THĂM ĐỒ



KẾT QUẢ PHÂN TÍCH



KẾT LUẬN

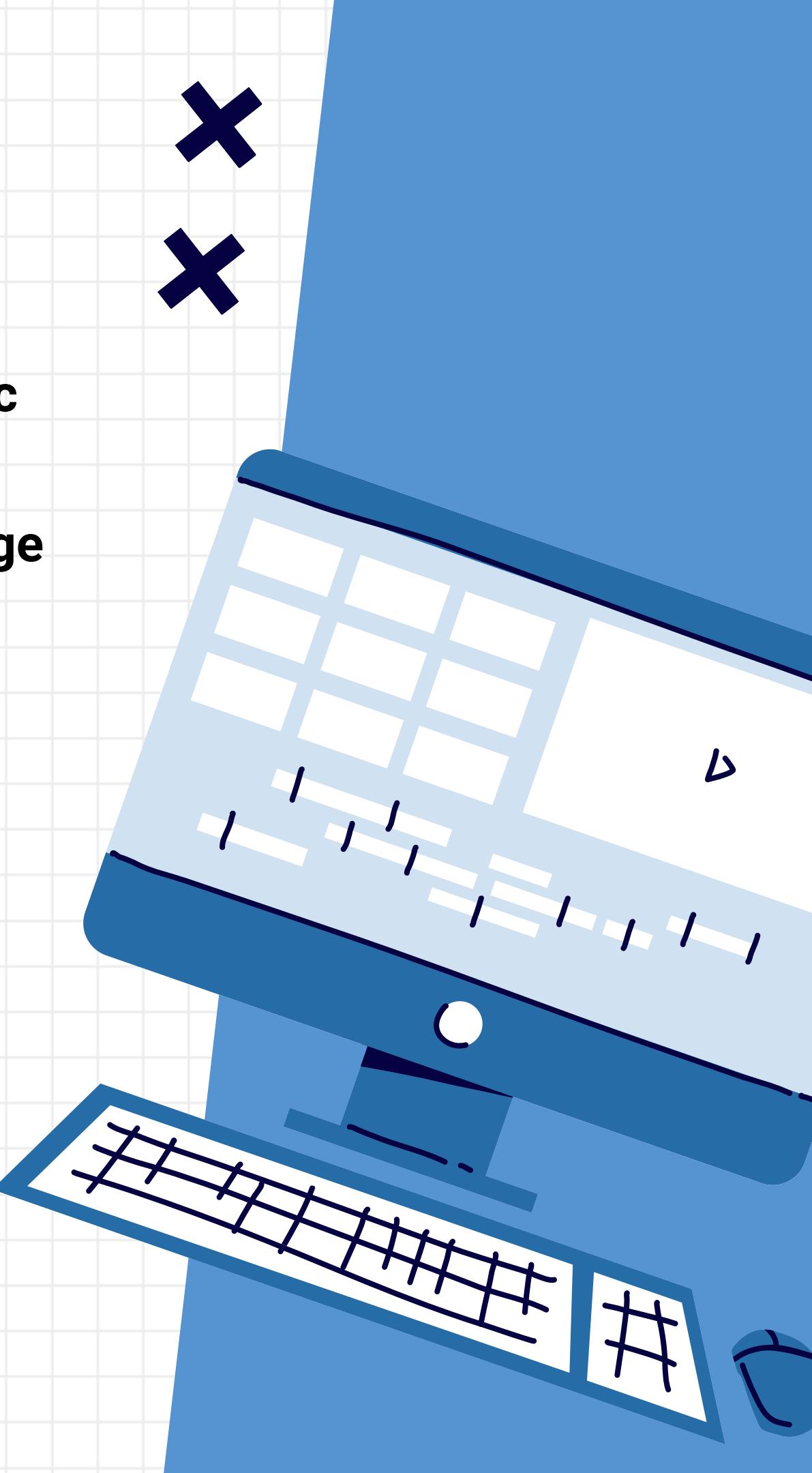
GIỚI THIỆU

Tầm quan trọng

- Phân tích yếu tố ảnh hưởng tới doanh thu phim giúp tối ưu hóa chiến lược kinh doanh, phân bổ ngân sách, và xác định yếu tố thành công.
- Áp dụng các mô hình học máy: LightGBM, Random Forest, XGBoost, Ridge Regression, Linear Regression.

Dữ liệu sử dụng

- Nguồn dữ liệu:
- *imdb.csv* (tự thu thập, cập nhật tới 2023, 1000 dòng sử dụng Python BeautifulSoup).
- *keywordsmovie.csv* (Kaggle, cập nhật tới 2017, 45465 dòng).
- Nhóm tự thiết kế, chỉnh sửa, và thêm tính năng để phục vụ nghiên cứu.
- Cam kết minh bạch về nguồn gốc dữ liệu và sẵn sàng cập nhật cải tiến.



MÔ TẢ BỘ DỮ LIỆU

Các cột chính trong *imdb.csv*

- Thông tin phim: Name of movie, Year of release, Watchtime, Genre, Classification.
- Đánh giá & doanh thu: Movie Rating, Metascore, Votes, Gross collection.
- Mô tả: Des.

Các cột trong *keywordsmovie.csv*

- Title: Tiêu đề phim (khớp với Name of movie trong *imdb.csv*).
- Keywords: Từ khóa mô tả nội dung phim.

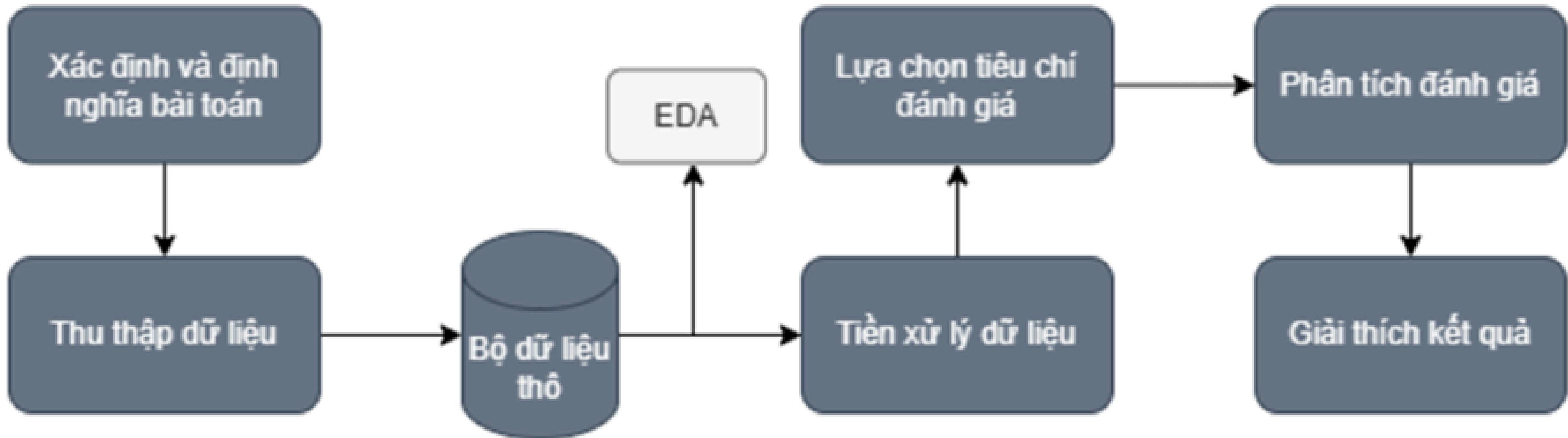
Liên kết dữ liệu

- Cột Name of movie là khóa ngoại, giúp gộp 2 bộ dữ liệu thành *movieofficial.csv* với 1,000 bộ phim và các cột mở rộng gồm "Keywords".



Cột dữ liệu	Giá trị	Kiểu dữ liệu
Name of movie	Iron Man 3	Object
Year of release	2013	int64
Watchtime	130	int64
Genre	Action, Adventure, Sci-Fi	Object
Classification	PG-13	Object
Movie Rating	7.1	float64
Metascore	62	float64
Votes	896,434	Int32
Gross collection	1,215,577,205	float64
Des	When Tony Stark's world is torn apart by a formidable terrorist called the Mandarin, he starts an odyssey of rebuilding and retribution.	Object
Keywords	terrorist, war on terror, tennessee, malibu, marvel comic, superhero, based on comic, tony stark, iron man, aftercreditsstinger, marvel cinematic universe, mandarin, 3d, war machine, iron patriot, extremis	Object

QUY TRÌNH THỰC HIỆN



Làm sạch

Chuyển đổi kiểu dữ liệu

Xử lý giá trị thiếu

Giảm chiều

Giảm chiều: 100 chiều xuống 20 chiều.

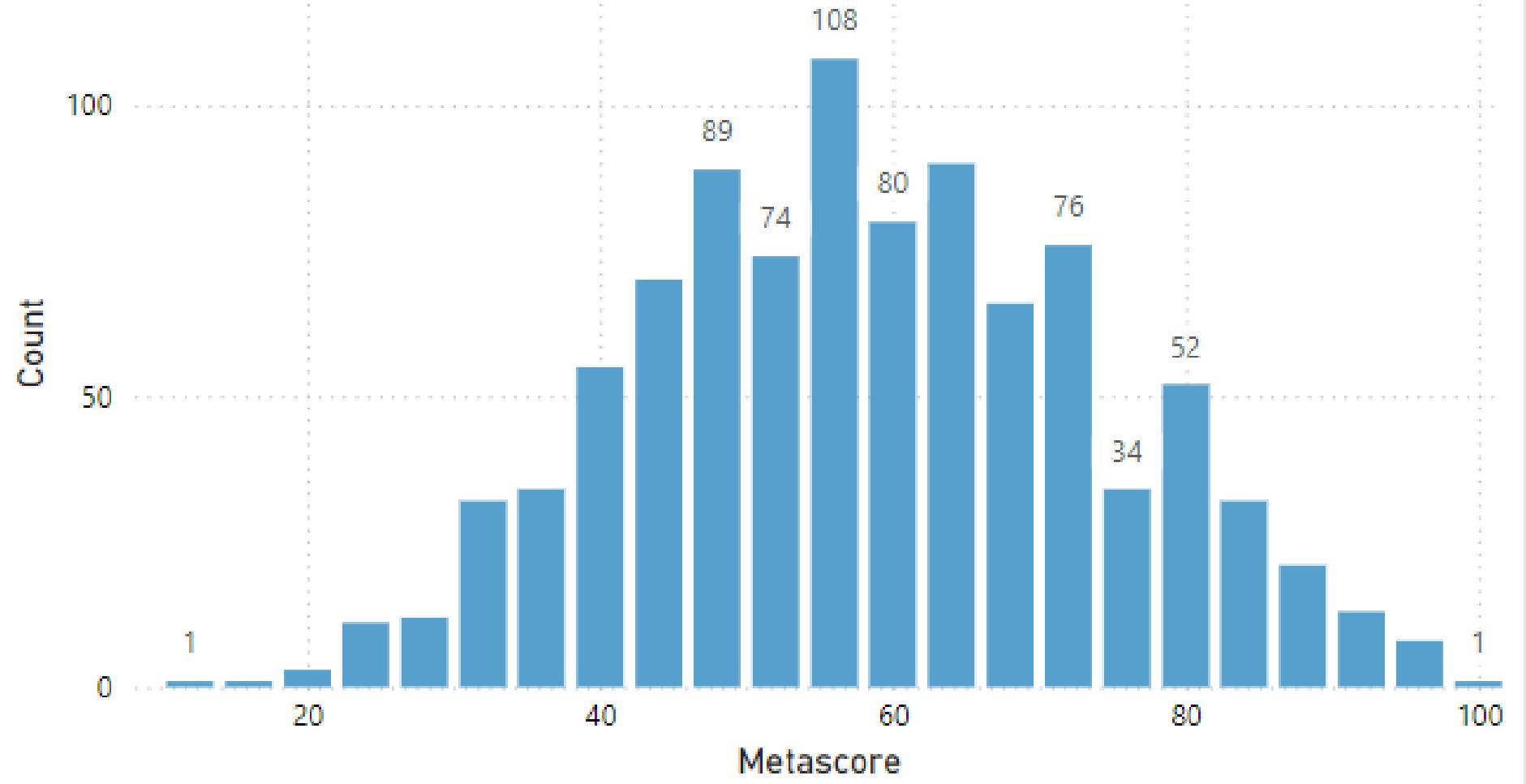
Mã hóa

Mã hóa từ khóa: chọn 100 từ khóa quan trọng nhất

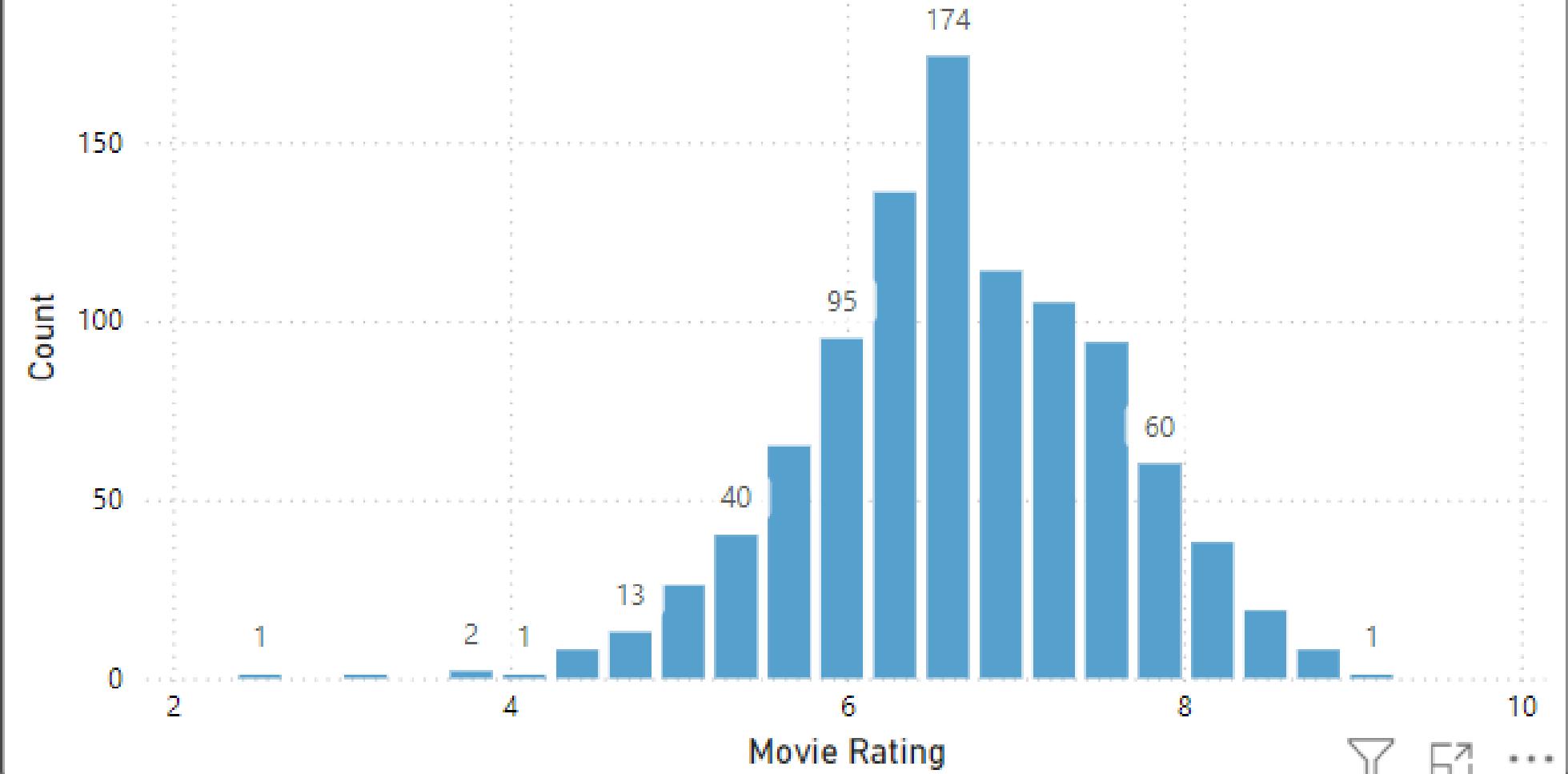
Chuẩn hóa

chuẩn hóa (scaling) để đưa các giá trị về một thang đo chung

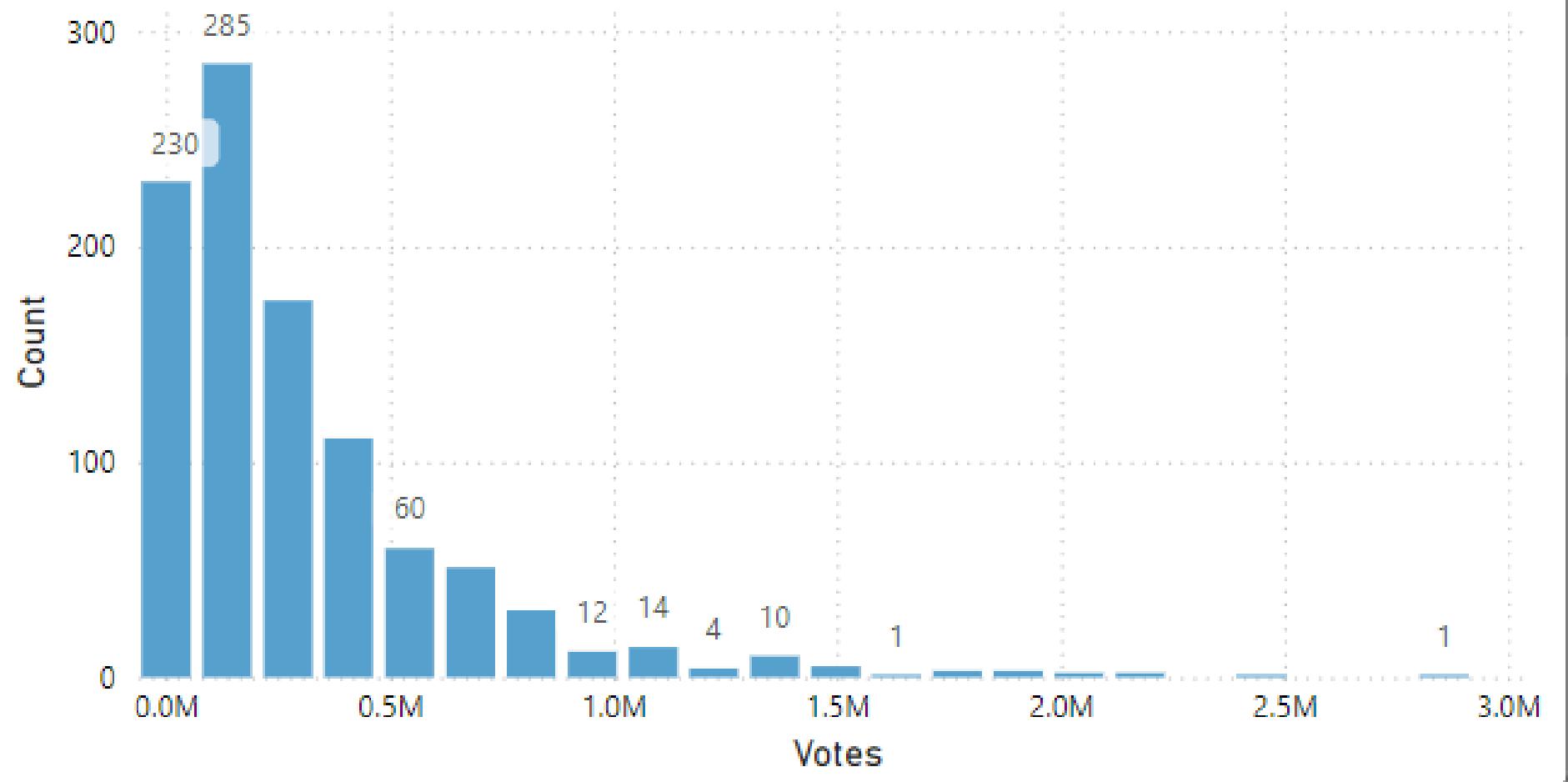
HÌNH 1: Phân phối Metascore



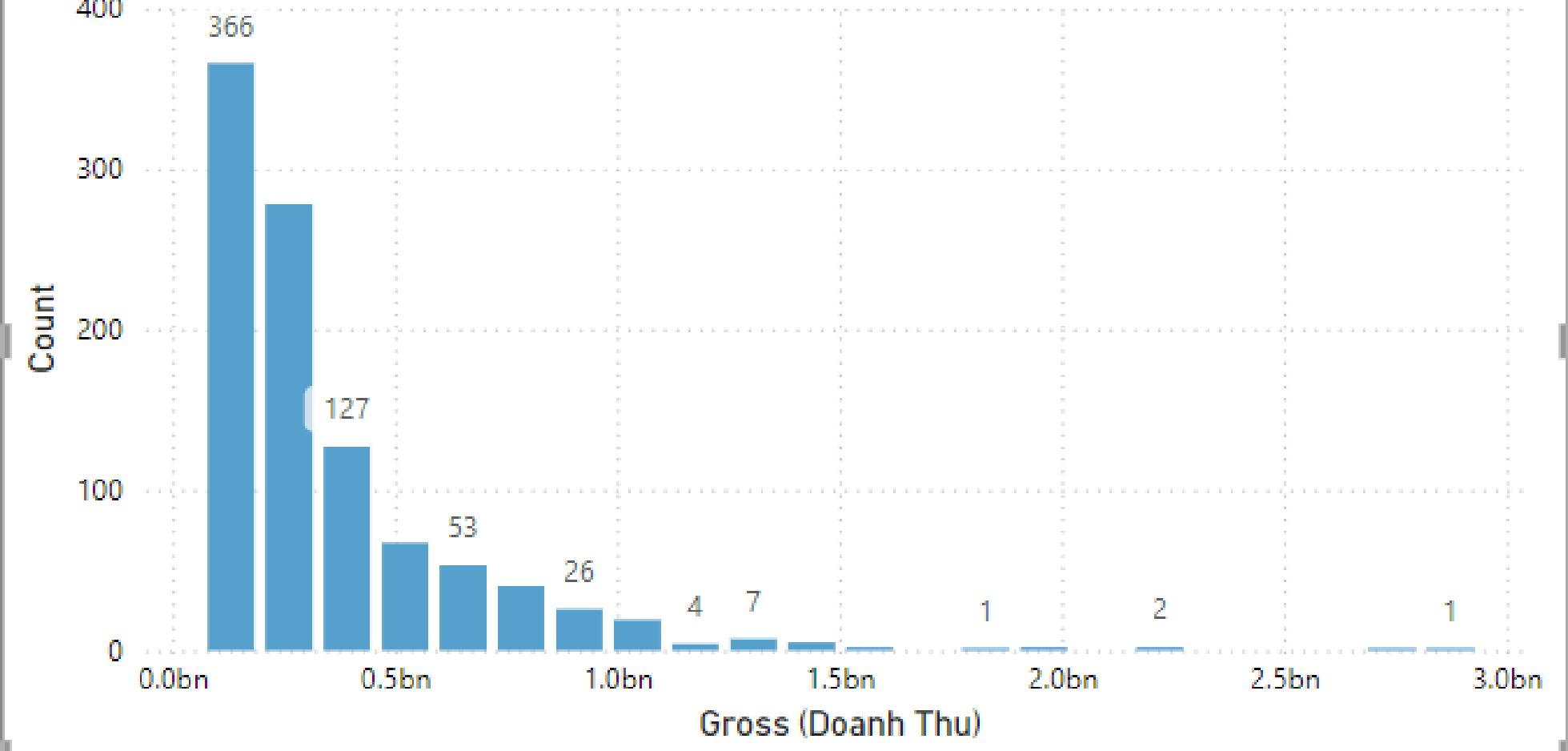
HÌNH 2: Phân phối Rating



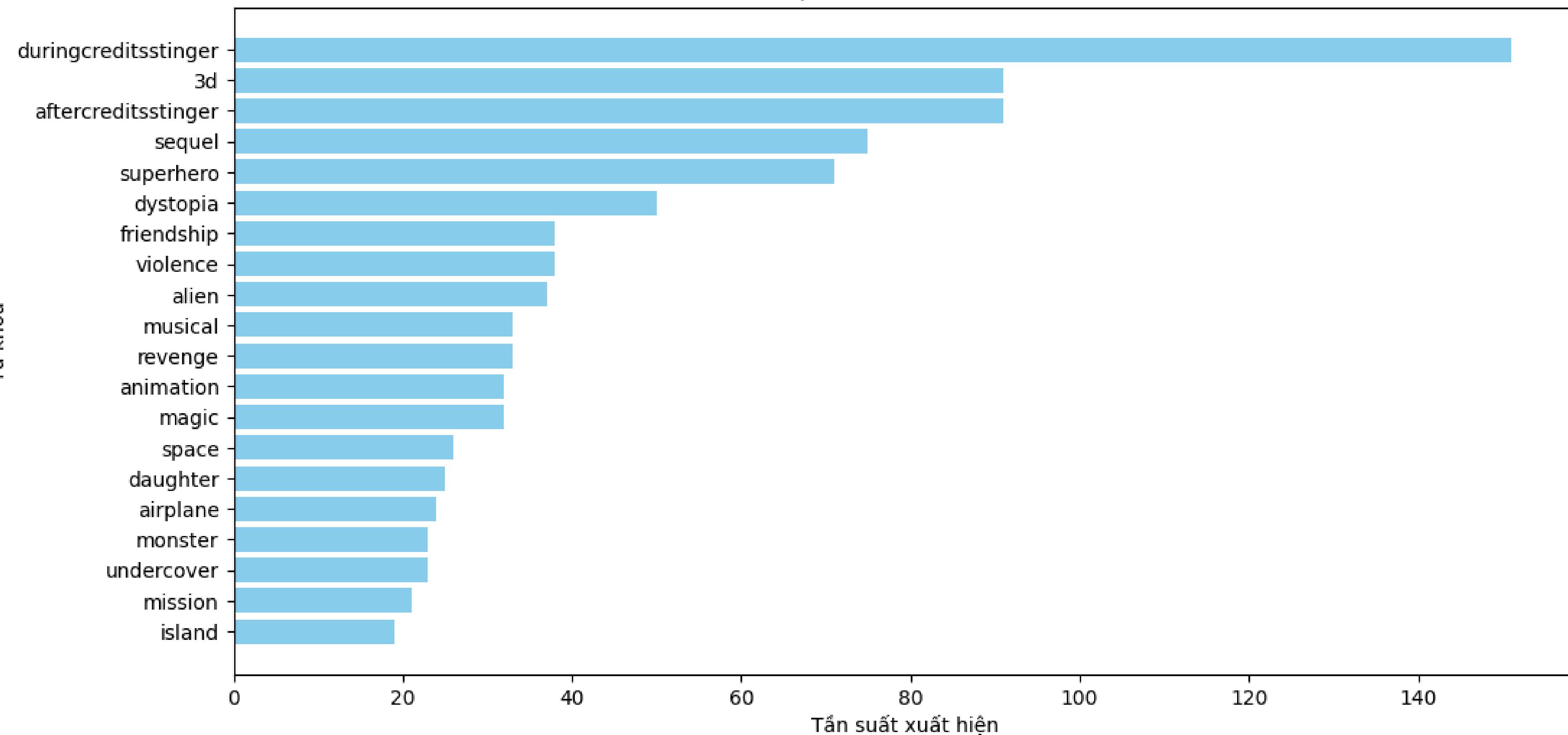
HÌNH 3: Phân phối Votes



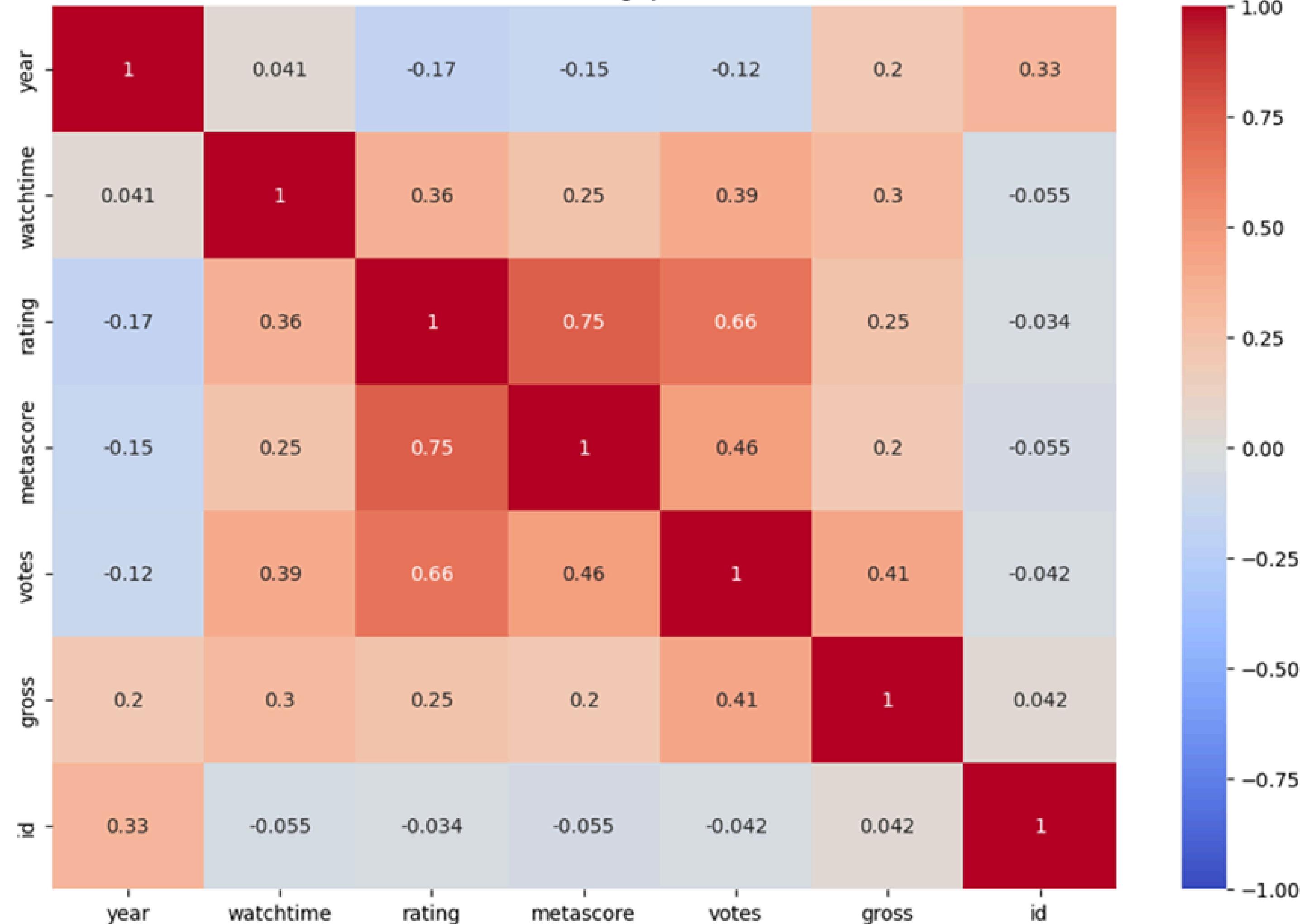
HÌNH 4: Phân phối Doanh Thu



HÌNH 5: Top 20 Từ Khóa Phổ Biến Nhất



Ma trận tương quan



THIẾT KẾ TIÊU CHÍ ĐÁNH GIÁ



BIỂN

- Metascore
- Rating
- Votes
- Keywords
- VS DOANH THU

MÔ HÌNH VÀ ĐÁNH GIÁ

- Random Forest Regressor
- XGBoost Regressor
- LightGBM Regressor
- Ridge Regression
- Linear Regression

- MSE (Mean Squared Error)
- R² Score
- MAE (Mean Absolute Error)

TIÊU CHÍ

- MSE, MAE càng thấp là tốt nhất, R² Score càng gần 1 là tốt nhất.
- dựa theo độ lệch giá trị của DOANH THU

KẾT QUẢ

Model	R ²	MSE	MAE
Linear Regression	0.198366	7.356812e+16	1.884538e+08
Ridge Regression	0.200717	7.335239e+16	1.881428e+08
Random Forest	0.160791	7.701649e+16	1.845953e+08
XGBoost	0.098822	8.270359e+16	1.935465e+08
LightGBM	0.246584	6.914303e+16	1.809380e+08

- LightGBM: Là mô hình hiệu quả nhất sai số thấp nhất. (*)
- Ridge Regression và Linear Regression: Cho kết quả ổn định
- Random Forest và XGBoost hiệu suất thấp hơn với sai số lớn.

LightGBM



Chỉ số tầm quan trọng (importance) của các biến là thước đo phản ánh mức độ ảnh hưởng của mỗi biến (hoặc đặc trưng) trong việc dự đoán ảnh hưởng



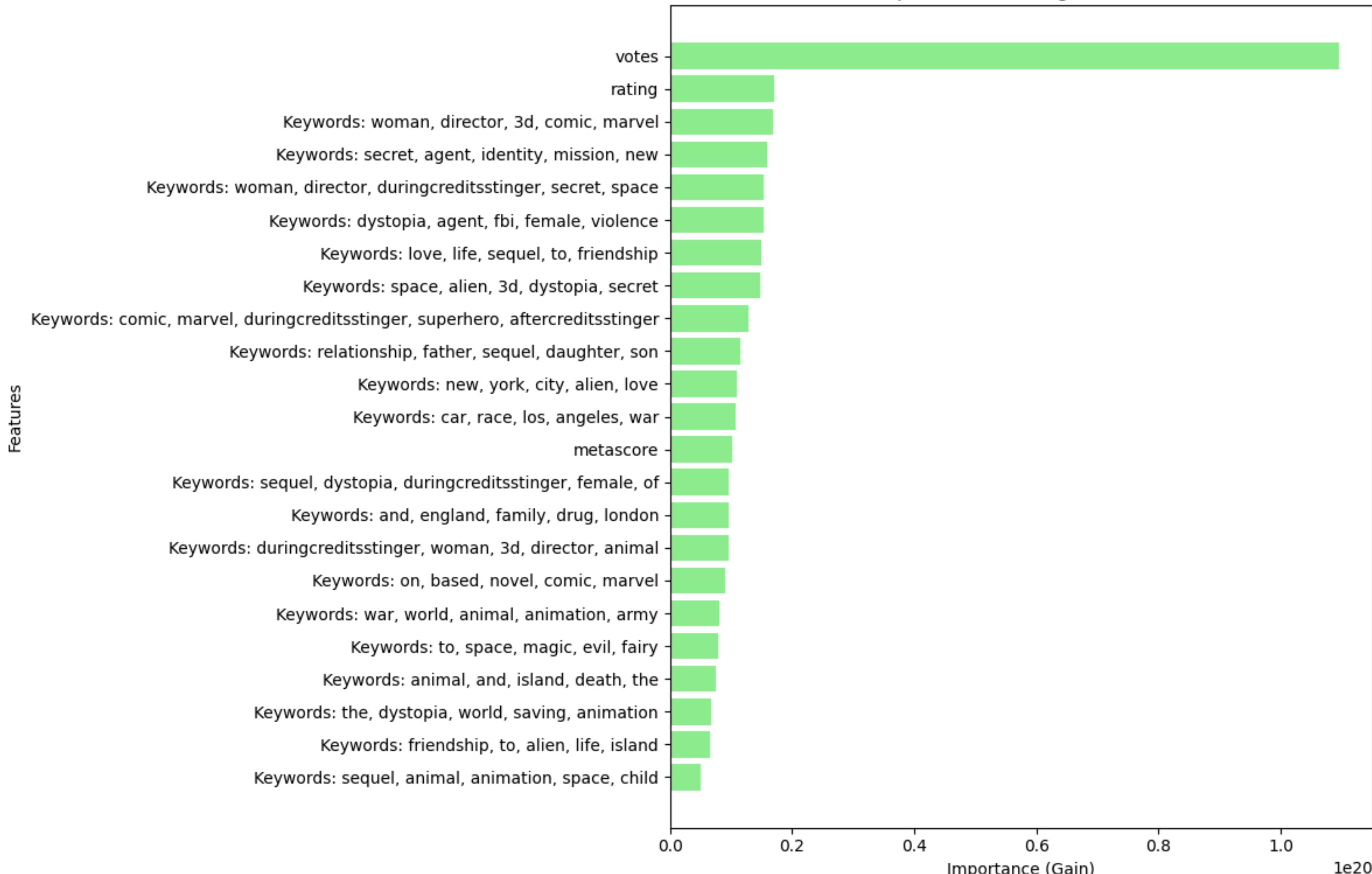
Chỉ số tầm quan trọng

- Weight
- Gain
- Cover



Gain được sử dụng làm chỉ số chính trong nghiên cứu này vì nó phản ánh mức độ quan trọng của từng đặc trưng trong việc giảm độ lỗi và cải thiện hiệu suất mô hình.

Feature Importance from LightGBM (With Gain)



KẾT LUẬN

Kết quả đạt được

- Xây dựng bộ dữ liệu và tiền xử lý hoàn chỉnh
- Xây dựng được các mô hình phân tích các yếu tố ảnh hưởng đến doanh thu.
- Tìm ra được tầm quan trọng các biến ảnh hưởng tới doanh thu tốt

Hạn chế

- Kết quả mô hình không được như kỳ vọng.
- Chưa tìm ra các “keywords” quan trọng đóng vai trò trong việc ảnh hưởng tới doanh thu phim.





THANK YOU