Appendix

Data Summary

This dataset was created by merging the ACL IMDB Movie Reviews and IMDb titles basic where each row in the dataset represents a single IMDb movie review. The text review and number of review stars is linked to an official movie identifier from IMDb (tt######) and merged with IMDB metadata including the movie's title and genres. However, since some movies belong to multiple genres, a single review may appear multiple times in the dataset, once per genre. Each row represents a unique title in the IMDb catalog along with its IMDb identifier, release year, and genres. This final merged dataset has a total of 114137 observations.

Provenance

The dataset used for this project is a corpus of movies based on an IMDb dataset created by Andrew Maas and his colleagues at Standford. It consists of 50,000 reviews from IMDb's database, originally divided into 25,000 training and 25,00 for testing, each evenly split between positive and negative reviews. We combined these into a single collection to apply our own splits during the model building process. Each review is linked to an associated IMDb movie identifier which allowed us to merge the review texts with the official IMDb metadata consisting of the movie's name and genre information.

License

MIT License

Copyright (c) 2025 ashleynguyen04

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Ethical Statements

The data used for this project, ACL IMDB Movie Review Dataset and the IMDb Title Basics file, are both publicly available and intended for academic purposes.

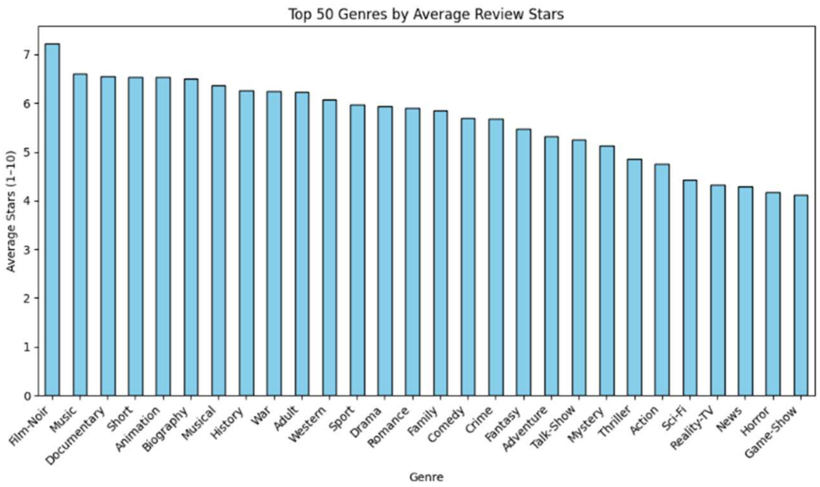
The ACL IMDB corpus consists of movie reviews originally posted publicly on the IMDb website. While no personally identifying information is included in the dataset, as files only include the review text, number of review stars, and associated movie identifier, it is important to recognize that the reviews were written by real users. The dataset was anonymized by researchers (Maas et al., 2011) to minimize privacy risks.

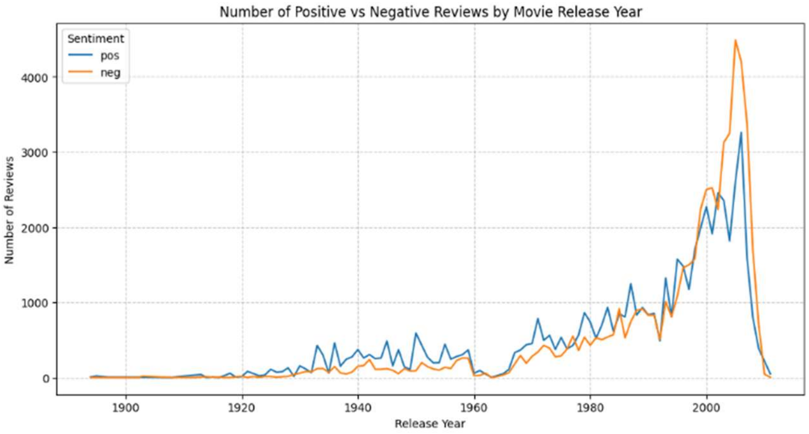
The datasets used here are strictly for educational and research purposes—specifically for exploring sentiment analysis and genre-based trends. The analysis does not attempt to deanonymize users or commercialize IMDb's data.

Data Dictionary

Field	Type	Description	Example
tt_id	string	IMDb identifier	tt0100680
movie_name	string	Full title of movie	Stanley and Iris
genres	string	List of genres if movie falls into multiple categories	Drama, Romance
genre	string	Drama	Drama
review	string		"This isn't the comedic Robin Williams, nor is it the quirky/insane Robin Williams of recent thriller fame. This is a hybrid of the classic drama without over-dramatization, mixed with Robin's new love of the thriller…."
rating	string	Whether the review was positive (≥7) or negative (≤4)	pos
review_stars	integer	Number of stars reviewer gave movie (scale from 0-10)	3

Exploratory Plots


Top 50 Genres by Average Review Stars
Average Stars (1-10)
Film-Noir
Music
Documentary
Short
Animation
Biography
Musical
History
War
Adult
Western
Sport
Drama
Romance
Family
Comedy
Crime
Fantasy
Adventure
Talk-Show
Mystery
Thriller
Action
Sci-Fi
Reality-TV
News
Horror
Game-Show
Genre


Number of Positive vs Negative Reviews by Movie Release Year
Sentiment
pos
neg
Number of Reviews
Release Year