Appendix

ACL IMDb Movie Reviews Dataset

The raw reviews dataset has the reviews stored as text files inside train/pos, train/neg, test/pos, and test/neg folders. Each file contains the full review the and the filename corresponds with the user's star rating (0-10). The associated IMDb movie identifier (tt#####) is in a separate folder called urls_pos.txt and urls_neg.txt.

Movie Review Attributes

- Review (raw text): full text of a user's movie review
- Rating (binary): encoded by folder location in "pos" meaning reviewer left ≥7 stars or "neg" meaning reviewer left ≤ 4 stars
- Review_stars (1-10): encoded by review file name
- Imdb_identifier (tt#######): extracted for urls_pos,.txt or ursl_neg.txt files

Descriptive Statistics

- Total reviews: 50,000
- Review rating distribution: 25,000 positive and 25,000 negative
- Average review length: ~230 words

IMDb Title Basics Dataset

Each row represents a unique title in the IMDb catalog along with its IMDb identifier, release year, and genres. For the final dataset creation, only the IMDb identifiers found in the ACL IMDb Movie Reviews Dataset were retained.

Variables

Field	**Type**	**Description**	**Example**
tconst	string	Unique IMDb identifier	tt0100680
primaryTitle	string	Full title of movie	Stanley and Iris
startYear	integer	Year of initial release	1990
genres	list	comma-separated listed of genres	Drama, Romance

Descriptive Statistics

- Total observations: 9 million rows (including movie, short, or TV show)
- Top genres: Drama, Comedy, Thriller, and Action

IMDb Sentiment with Genres Dataset

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.

Variables

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

Descriptive Statistics

- Total observations: 114137 rows
- Unique movies: ~5,000