Classifying IMBd Reviewsn (MI1)

EAA Sports: Ashley Nguyen (leader), Emma Wunderly, and Andre Galistinos

DS 4002-001 | 9/12/25

Our project is centered on movies and their reviews. We plan to use the text data from IMDb to classify reviews on a scale of positive and negative based on keywords in the text. By analyzing these reviews, we hope to discover which genres have the most positive and negative reviews, serving the movie industry and consumers with a gauge on reactions and therefore what can be done better within genres.

Our target audience is those working in the movie industry and movie watchers. We find that movie reviews are crucial in the film industry as they provide filmmakers with an idea of who their audience is, to continue making movies that are better than the last. But it also gives consumers a sense of what a movie is about and whether it is worth watching. We plan to create a model that defines both positive and negative words to classify reviews. The steps we will take to achieve this begin with using a model that can predict movie review sentiment. We will then classify these reviews based on genre to reveal which genre categories fall under the most and least positive and negative reviews.

* Build a model that can predict sentiment
* Making a trained set to be used for our test set
* Then classifying based on the genre

the Self Moitoring Scale defined as: “Self-observation and self-control guided by situational cues to social appropriateness” (Snyder, 1974).

<https://scholars.unh.edu/cgi/viewcontent.cgi?params=/context/honors/article/1267/&path_info=Jacob_Pentheny_Honors_Thesis_Influence_of_Movie_Reviews.pdf>

<https://libguides.unm.edu/film/reviews-criticism>

Goal statement

* The goal of our project is to develop a model that classifies IMDb movie review as positive or negative allowing us to easily summarize audience sentiment.

Research question

* Do certain movie genres receive a more negative or positive reaction by reviewers?

Modeling approach (1 paragraph)

For this project, our sentiment analysis for movie reviews would start with baseline models such as logistic regression or Naive Bayes. These would be highly interpretable and helps us understand which features, such as words or phrases, are most strongly associated with positive or negative reviews. We will preprocess the data and create a labeled tarining data set to They will serve as reliable benchmarks and have been shown in previous research to perform well in text classification tasks (1.) Next, we could implement random forests which could be effective in non-linear features interactions and could improve classification performance (3)

(1)<https://www.researchgate.net/publication/378691218_Sentiment_Analysis_of_IMDb_Movie_Reviews_Using_Traditional_Machine_Learning_Techniques_and_Transformers>

<https://rpubs.com/knlussi/draft1117>

(3) <https://www.deanfrancispress.com/index.php/te/article/view/920>

* Baseline model: Logistic regression
* Random Forest

References (All)



Colab: <https://colab.research.google.com/drive/1BpahIHacAxGgjNVOfgPh7_QEKWnawOmS?usp=sharing>

Github: <https://github.com/ewunder32/Project_1>

Datasets: [IMDb Data Files Download](https://datasets.imdbws.com/)

[Sentiment Analysis](https://ai.stanford.edu/~amaas/data/sentiment/)

Alien, song -