**Performing Sentiment Analysis on IMDB Reviews - Market Research for Future Movie Productions**

**Team Members:**

Richard Anderson, Min Chang, Sharmishta Kumaresh

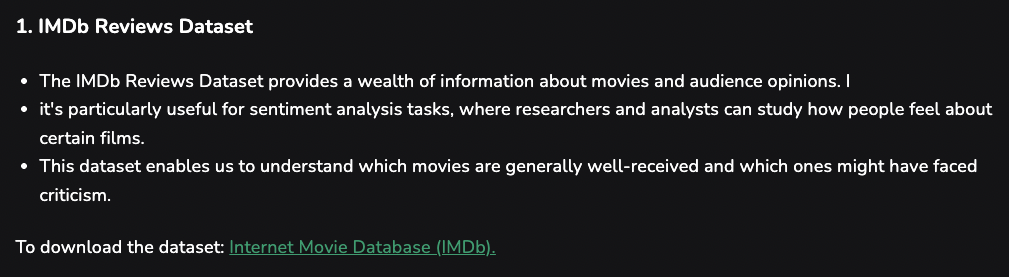
**Problem Statement:**

Movie production companies need relevant information on how movies are received by their audiences. Online reviews provide a real-time view into audience reactions and are one of the fastest ways to identify how a movie will do at the box office. Additionally, audience reactions can predict future trends and help production companies identify what projects to pursue.

In this study, we will be completing domain-specific sentiment analysis on a dataset of Internet Movie Database (IMDB) movie reviews. Our goal is to identify if the reviews are positive or negative. This rating can be used by movie production companies to identify movie preferences for future productions.

**Dataset Selection:**

The dataset was sourced from the following link: <https://www.geeksforgeeks.org/dataset-for-sentiment-analysis/>



This dataset consists of approximately 50,000 movie reviews. Each review has already been categorized as positive or negative. This dataset is a csv file with two columns. The first is the review and the second is the categorization.

**Expected Outcomes:**

For this project, our plan is to identify which movies performed well and those that did not by applying sentiment analysis on the reviews. This will help us gauge what aspects of these movies resonated with or polarized crowds. As the reviews have already been categorized as positive or negative, we have true values to compare the results of our analysis with. In addition to building a sentiment classifier, we also want to explore the patterns in the text itself. By analyzing the language used in both positive and negative reviews, we hope to see which words, phrases, or topics come up most often for each sentiment. For example, we might find certain actors, movie genres, or specific types of storytelling are mentioned more in positive reviews while others might appear more in negative reviews. Overall, our biggest expectation from this project is to apply NLP and ML techniques to real-world text data, while gaining hands-on experience exploring different models and approaches for text analysis.

**Methodologies:**

In order to achieve the objectives of our project, we will be:

* Loading the IMDB dataset into a pandas DataFrame to check for any issues such as duplicates, missing values, or any unexpected characters.
* Cleaning the text data by removing any HTML tags and special characters, and converting the entire text data into lowercase for consistency.
* Applying Named Entity Recognition (NER) to extract key entities such as actor names, locations, or movie titles, analyzing how these entities relate to sentiment patterns.
* Tokenizing the reviews by splitting each text into a list of words and converting text into numerical features using vectorization techniques (TF-IDF/Count Vectorization)
* Splitting the data into training and testing sets in order for our model to evaluate on unseen data (possibly 80% training and 20% testing)
* Training our ML model by using a few baseline models such as Logistic Regression, Random Forest, and/or support vector machines.
* We may also use Neural Network Models such as Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) networks to grab sequential patterns from the IMDB text reviews.
* Model Optimization (Adam, Quasi-Newton methods, loss functions, gradient descent) + Hyperparameter tuning
* ML Model Result Evaluation: comparing our predictions against the true positive/negative labels using metrics like precision, accuracy, recall, and F1-Score
* Visualization
* Error Analysis