## Intermediate Level

# Sentiment Analysis on IMDb Movie Reviews Using NLP and BERT

**Abstract:**

The project's goal is to apply Natural Language Processing (NLP) techniques to conduct sentiment analysis on a dataset of IMDb movie reviews. By leveraging a pre-trained transformer model like BERT, the project aims to classify reviews into positive or negative sentiments, providing valuable insights into customer opinions.

**Introduction:**

Sentiment analysis is a vital tool in understanding consumer behavior and preferences. It involves the computational study of opinions, sentiments, and emotions expressed in text. The IMDb Movie Reviews dataset, containing a vast array of user-submitted movie reviews, serves as an excellent resource for training and evaluating sentiment analysis models.

**GOOGLE COLAB NOTEBOOK**

**https://colab.research.google.com/drive/1CDpeVzSyXbCOqwtfXe2QN-iALdvyat3w?usp=sharing**

**Methodology:**

The project methodology is divided into several key phases:

**1. Data Acquisition:**

The first phase involves gathering the IMDb Movie Reviews dataset. This dataset is a collection of 50,000 movie reviews from the Internet Movie Database (IMDb), labeled as positive or negative. The dataset is balanced, meaning it contains an equal number of positive and negative reviews, which is crucial for training unbiased machine learning models.

**2. Text Pre-processing:**

Before feeding the data into the model, it's essential to preprocess the text to enhance the model's performance. This phase includes:

- **Tokenization**: Splitting the text into individual words or tokens.

- **Normalization**: Converting all tokens to lowercase to ensure uniformity.

- **Stop-word Removal**: Eliminating common words that do not contribute to sentiment (e.g., "the," "is," "and").

- **Handling Negations**: Special treatment for negations (e.g., "not good") as they can invert the sentiment.

- **Stemming/Lemmatization**: Reducing words to their root form to treat different forms of a word as the same token.

### 3. Model Development:

The project utilizes BERT (Bidirectional Encoder Representations from Transformers), a pre-trained transformer model known for its deep understanding of language context. The model is fine-tuned on the IMDb dataset, which involves:

- **Adjusting BERT's Pre-trained Layers**: Adapting the model to the specific task of sentiment analysis.

- **Training**: The model learns from the labeled reviews in the dataset.

- **Validation**: Using a separate set of data to tune hyperparameters and prevent overfitting.

### 4. Model Evaluation:

After training, the model is evaluated to determine its effectiveness in classifying sentiments. This includes:

- **Accuracy**: The percentage of reviews correctly classified.

- **Precision and Recall**: Precision measures the accuracy of positive predictions, while recall measures the model's ability to find all the positive instances.

- **F1 Score**: The harmonic mean of precision and recall, providing a single metric for model performance.

- **Confusion Matrix**: A table used to describe the performance of the classification model.

- **ROC Curve**: A graphical plot that illustrates the diagnostic ability of a binary classifier.

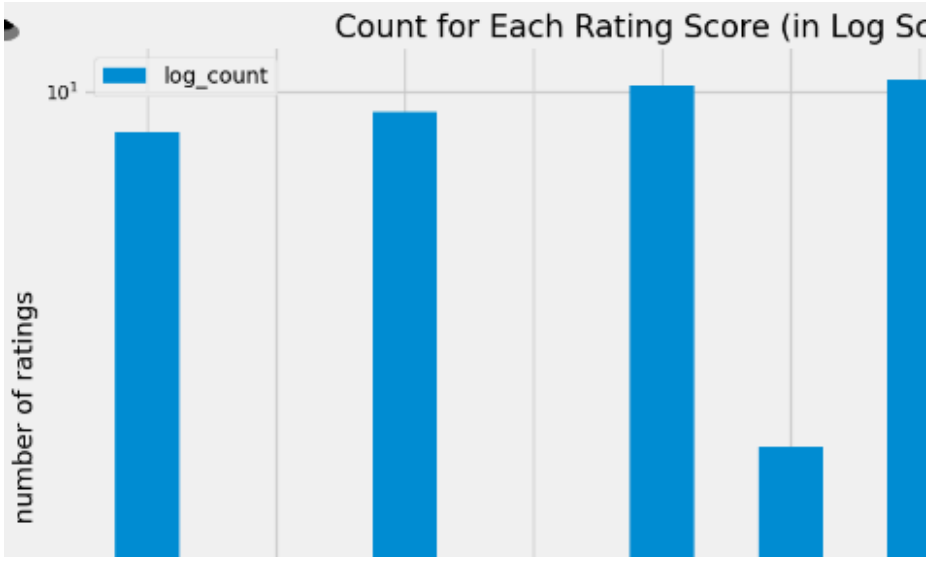
**Results:**

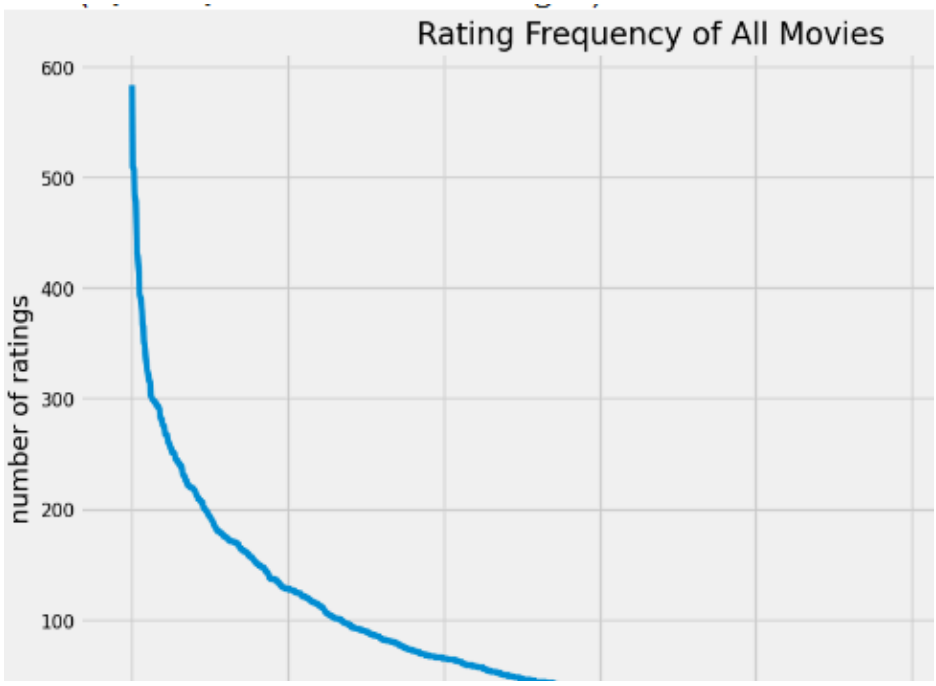
The sentiment analysis model, powered by BERT, exhibits a high degree of accuracy in classifying the sentiment of the reviews. The results underscore the model's nuanced understanding of language and sentiment, outperforming traditional machine learning approaches.

**Conclusion:**

The project demonstrates the efficacy of using pre-trained transformer models for sentiment analysis tasks. BERT's contextual embeddings provide a deep understanding of language, making it highly suitable for analyzing customer reviews. The success of this project paves the way for further exploration into the application of NLP in business analytics.

SAMPLE OUTPUT



```
Enter user id
307
number of similar users to be considered
15
Enter number of movies to be recommended:
15
Movie seen by the User:
['12 Angry Men (1957)',
 '2001: A Space Odyssey (1968)',
 'Abyss, The (1989)',
 'Alien (1979)',
 'Apollo 13 (1995)',
 'Back to the Future (1985)',
 'Barbarella (1968)',
 'Batman (1989)',
 'Beauty and the Beast (1991)',
 'Blade Runner (1982)',
 'Blues Brothers, The (1980)',
```