Assignment 2

Name: Mohamed-Obay Alshaer
Student Number: 300170489

Submission Date: March 21, 2025

Professor: Daniel Shpairo

SEG4300 - Applied Machine Learning

1 IMDb Reviews Dataset Exploratory Data Analysis

1.1 Dataset Selection

The IMDb Reviews Dataset (stanfordnlp/imdb) was selected for this analysis. This dataset contains 50,000 movie reviews evenly divided between training and testing sets, with each review labeled as positive (1) or negative (0). The dataset was chosen for its relevance as a benchmark for sentiment analysis, manageable size, quality documentation, balanced class distribution, and rich text content suitable for various NLP applications. For computational efficiency, a random sample of 10,000 reviews was used while maintaining dataset representativeness.

1.2 Key Findings

- **Data Quality:** The dataset is well-balanced with no missing values in core fields. All labels are valid (0 or 1). A small number of extremely short reviews and some duplicates were identified.
- **Text Characteristics:** Review lengths vary significantly (from under 100 to over 10,000 characters), with a median length of approximately 800 characters or 150 words. Negative reviews tend to be slightly longer than positive ones.
- **Content Analysis:** Clear vocabulary patterns distinguish positive and negative reviews. Positive reviews frequently contain words like "great", "excellent", and "best", while negative reviews commonly include "bad", "worst", and "boring".

1.3 Recommendations for Modeling

- **Text Preprocessing:** Normalize review lengths, implement thorough text cleaning, and apply stemming or lemmatization.
- **Feature Engineering:** Explore n-gram features, consider TF-IDF weighting, and extract additional features like review length and punctuation patterns.
- **Model Selection:** Test traditional approaches (Naive Bayes, SVM) as baselines and consider deep learning approaches (BERT, RoBERTa) for advanced performance.
- **Evaluation Strategy:** Use stratified cross-validation, focus on comprehensive metrics (precision, recall, F1-score), and analyze error cases.

1.4 Conclusion

The IMDb Reviews Dataset provides a rich source for sentiment analysis with clear linguistic patterns between positive and negative reviews. The identified data quality issues are minimal and addressable during preprocessing. Even simple models may achieve reasonable performance on this well-structured dataset, while more sophisticated approaches could potentially achieve state-of-the-art results.