Project V: Sentiment Analysis for Movie Reviews

Scenario

You and your team are part of an analyst division in an organization that aggregates and analyzes user-generated movie reviews from various platforms. Your organization strives to become the leading hub for audience insights on movie content. Studios and streaming platforms invest heavily in movie production and marketing, but traditional box office numbers or star ratings do not necessarily reveal the underlying emotions, themes, or audience reception nuances. Therefore, the organization aims to harness natural language data to uncover viewer sentiment in a scalable, automated way. With a huge number of reviews posted daily, manual analysis is impossible. Accordingly, a robust sentiment analysis pipeline is needed to extract actionable insights from raw text reviews. Consequently, your team is tasked with developing a machine learning-based sentiment analysis to automatically classify reviews as positive or negative, allowing to provide sentiment dashboards for movies. For this purpose, your organization is equipped with both labeled and unlabeled user review data for various movies.

The board strongly desires that you adopt the proposed process model for project execution to ensure a systematic, transparent, and reproducible procedure for project management and technical implementation. Apart from this, you have no requirements on how to compose your technology stack (e.g., use of on-premise or cloud environments, tool selection for data storage, orchestration, analytics, etc.).

Data

You are provided movie review data from various platforms. In the following, the data schema is briefly described.

Attribute	**Description**
review	User review
label	Sentiment of the review, “Unknown” for the unlabeled data
movie_id	Unique identifier for the movie
reviewer_rating	Reviewer rating for particular movie (1-10), NaN for unlabeled data
movie_url	Movie URL for corresponding movie
title	Title of the movie

Table 1. reviews_labeled.csv/reviews_unlabeled.csv

Deliverables

- Extensive report containing:
    - Application of the concepts from the lecture
    - Documentation of all activities and components of the process model
    - Explanation of technology selection and design decisions
    - Deployment diagram(s) displaying the interconnection of the used tools
    - Outputs in context of the business scenario
- Code base