Data Requirements for Machine Learning Pipeline in MediaReview Social

Your Company Name

February 22, 2025

Contents

1 Introduction

This document outlines the data requirements for the machine learning pipeline of MediaReview Social. It describes the data sources, types, storage strategies, preprocessing, and data quality measures necessary to power our recommendation engine, sentiment analysis, and content moderation features.

2 Data Sources and Types

User-Generated Content

- **Reviews and Ratings:**
  - Text of reviews, numerical ratings (stars, thumbs up/down)
  - Metadata: timestamps, media identifiers (movie/show IDs), optional location data
- **Comments and Interactions:**
  - Replies, likes, shares, and threaded conversation data

User Profile and Behavioral Data

- **Profile Information:** Username, profile picture, bio, sign-up date
- **Usage Data:** Browsing history, clickstream data, session duration, frequency of interactions, and social connections (followers/following)

External Media Metadata

- **Media Information:** Cast, synopsis, genres, release dates, ratings (via APIs such as TMDb or IMDb)
- **Media Assets:** Posters, trailers, and other multimedia references

System Logs and Event Data

- **Interaction Logs:** API calls, user events (login, logout, posting), error events
- **Performance Metrics:** System and application metrics to monitor model performance and data flow efficiency

3 Data Volume, Velocity, and Variety

- **Volume:** Initially, thousands of reviews and interactions per day, scaling to tens of thousands as user adoption increases.
- **Velocity:** Combination of real-time ingestion for immediate features and batch processing for periodic model retraining.
- **Variety:** Structured data (user profiles, ratings) and unstructured data (review text, comments) requiring flexible storage and processing.

4 Data Storage and Pipeline

Raw Data Storage

- Store raw data in a data lake (e.g., AWS S3) using formats like JSON or CSV.

Processed Data Storage

- Structured data: Use relational databases (e.g., PostgreSQL).
- Unstructured data: Use NoSQL databases (e.g., MongoDB) or distributed systems (e.g., Apache Spark) for processing.

Data Pipeline Tools

- **Streaming Data:** Use Apache Kafka or similar platforms for real-time ingestion.
- **Batch Processing:** Employ tools like Apache Spark or AWS Glue for periodic processing and feature engineering.

5 Data Quality and Preprocessing

Cleaning and Validation

- Remove duplicates, correct inconsistencies, and filter out noise (e.g., spam or non-informative reviews).

Normalization and Transformation

- **Text Preprocessing:** Tokenization, lowercasing, and removal of stop words using NLP libraries (e.g., NLTK, spaCy).
- **Feature Engineering:** Convert text to embeddings (e.g., Word2Vec, TF-IDF, or transformer-based models) and extract behavioral metrics.

Handling Missing Data

- Establish strategies to impute or discard incomplete records to maintain data integrity.

6 Data Privacy and Security

- **Data Protection:** Encrypt data both in transit (TLS/SSL) and at rest.
- **Access Control:** Implement role-based access to restrict sensitive data.
- **Compliance:** Adhere to regulations like GDPR and CCPA, ensuring proper user consent and data anonymization.

7 Use Cases for Data in ML Models

- **Recommendation Engine:** Leverage user interaction and review history for personalized content suggestions.
- **Sentiment Analysis:** Analyze review text to determine sentiment, enhancing recommendations and highlighting trends.
- **Content Moderation:** Use user-generated content and logs to detect and flag spam or inappropriate content.