Exploratory Data Analysis (EDA) for Sentiment Analysis

This presentation will guide you through the process of conducting exploratory data analysis (EDA) for sentiment analysis, using the IMDB movie review dataset as an example.

Presented by

Maheshkumar Paik

Understanding the Dataset

Key Points

- 50,000 IMDB movie reviews categorized as **positive or negative**
- Dataset contains **text reviews** and corresponding sentiment labels

Identifying Class Imbalance

Key Points

Visualizing sentiment distribution

Key Points

Handling imbalanced data if necessary

Text Data Insights

Key Points

Word count & character count distributions

Key Points

How length affects sentiment

Word Cloud Analysis

Key Points

Most frequent words in IMDB reviews

Key Points

Removing stopwords to improve model training

Feature Selection - Correlation Matrix

Key Points

Analyzing correlation between word count, character count, and sentiment

Key Points

Selecting relevant features for ML models

Key Findings & Insights

Key Points

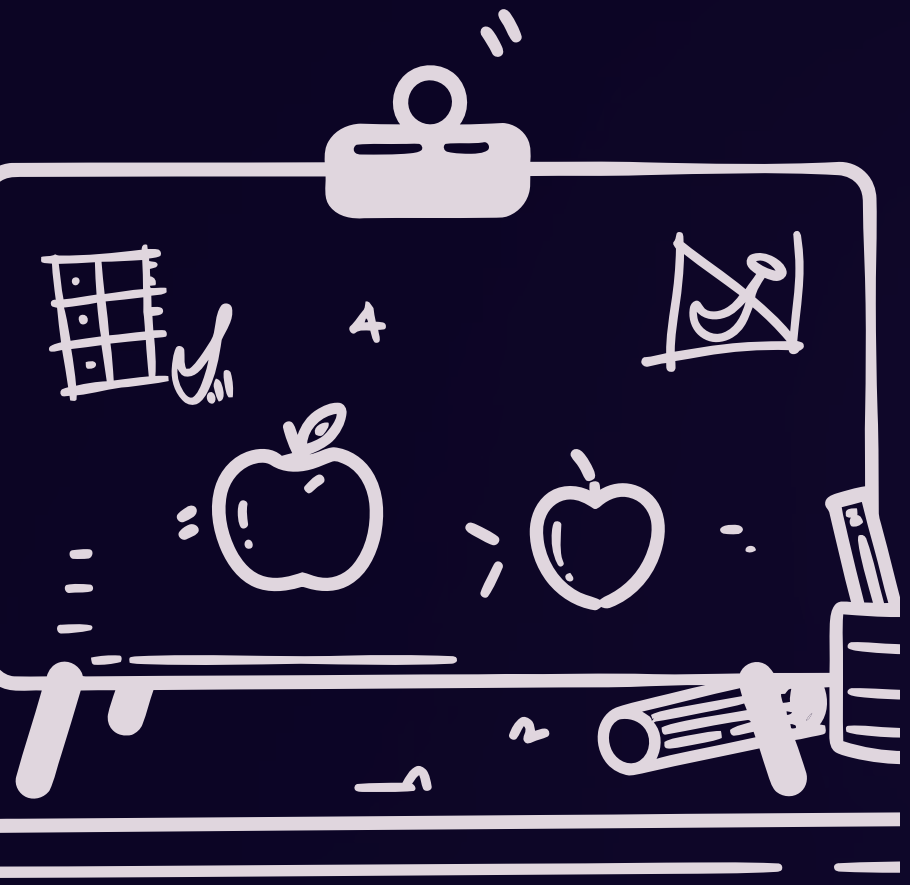
Balanced vs. imbalanced classes impact training

Key Points

Word count & character count may affect sentiment

Key Points

Common words reflect review patterns



Next Steps - Preparing for Machine Learning

Key Points

Use findings for feature engineering

Key Points

Train ML models for sentiment classification