

IMDB Movies Rating Sentiment Analysis Using Machine Learning Classifiers.

Abstract:

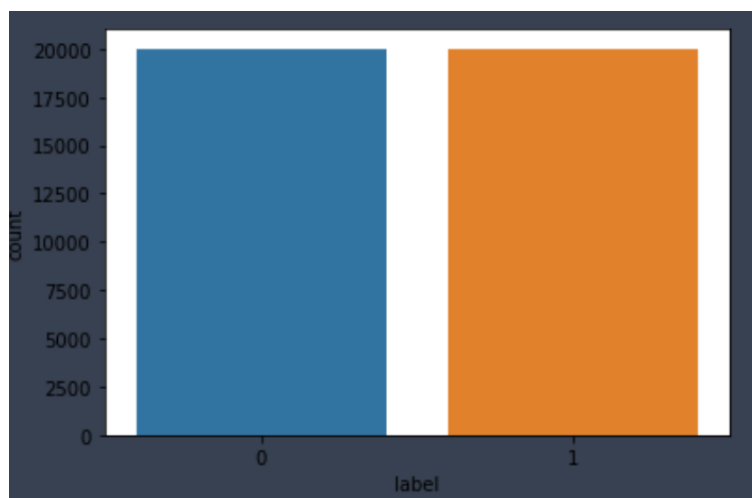
We are conducting research in this report. I extract a public dataset from Kaggle's *"IMDB Movies Rating Sentiment Analysis."* Data in the ".CSV" format received from a source is stale. To discover relevant predictions, I first performed certain data cleaning techniques, such as managing missing values and attribute validation and then apply a kNN classifier and TFIDF on it.

Introduction:

Content analysis, also known as information extraction, is the study of people's feelings, sentiments, assessments, attitudes, and emotions regarding things like products, services, organizations, persons, situations, events, themes, and their qualities. Sentiment analysis, in the context of movie review analysis, refers to determining how the general audience feels about a certain film. In this study, sentiment analysis tasks perform a binary categorization, such as analyzing positive and negative user evaluations.

IMDb is one of the most well-known internet databases for movies and celebrities, with millions of people reading and publishing movie reviews. Classifying the overall emotion polarity of movie reviews is a highly valued job. Sentiment analysis has long been a difficulty for the business, marketing, and management sectors in terms of adding value to the decision-making process.

Data Description:



The dataset contains forty thousand reviews of users and two classes **"class1 label 0"** represent negative reviews and **"class 2 label 1"** represent positive reviews of people. as shown in Figure 1. Binary class dataset.

Figure 1. Binary class dataset.

IMDB Movies Rating Sentiment Analysis Using Machine Learning Classifiers.

Methodology:

First of all, we apply cleaning tools to the dataset to achieve better accuracy. Then we are applying the kNN classifier with three neighbors to this data for sentiment analysis. And obtained the following results.

Training Accuracy 0.8663571428571428

0.7291666666666666

	precision	recall	f1-score	support
0	0.73	0.71	0.72	5970
1	0.72	0.75	0.73	6030
accuracy			0.73	12000
macro avg	0.73	0.73	0.73	12000
weighted avg	0.73	0.73	0.73	12000

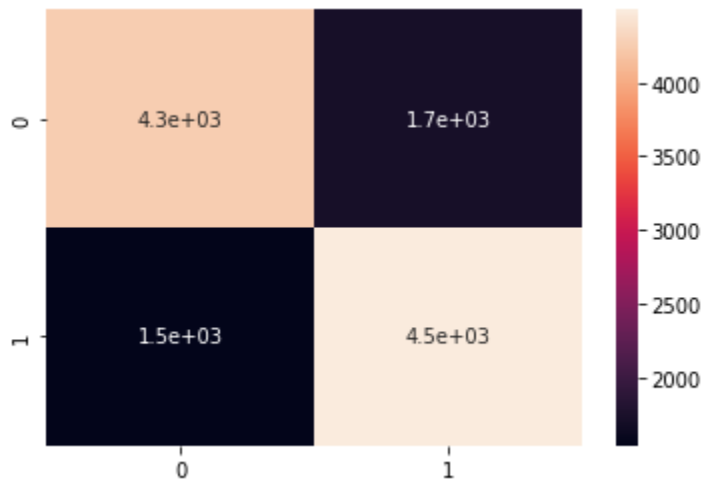


Figure 2. Confusion matrix

Conclusion:

In this research, we perform a simple analysis on a publicly available state-of-the-art "*IMDB Movies Rating Sentiment Analysis*." dataset. Two major tasks are performed in this research first one is data cleansing and the second wo using kNN n=3 classifier for the model training. We achieve remarkable accuracy no doubt its best accuracy is shown in **Error! Reference source not found.** but we improve it in the future.

Reference Dataset from Kaggle: <https://www.kaggle.com/datasets/yasserh/imdb-movie-ratings-sentiment-analysis>