The ML problem to solve

This project was built to solve a Supervision binary classification problem. Where the model determines if a movie review is negative or positive.

The dataset I used

I picked a dataset that contains 50k various movie review samples to train and test The model. So it can recognize differences between positive and negative reviews once it’s tested on unlabeled data.

DataSet: IMDB Dataset 50K: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Machine learning algorithm

I decided to go with Naive Bayes Multinomial, cause it is well suited for binary classfication problems.

What makes Naive Bayes a good alternative

- It is easy to implement to your application.
- Processing large amounts of data in short amout of time.
- Naive Bayes is well suited to solving text classfication problems, which make it a good pick for Sentiment Analysis

Optimize the model

- **clean the dataset, So I started removing unnecessary columns and symbols within the dataset to improve the performance of the model.**
- **There are a lot of common words in each review like: The, a, then and, many others. we can implement “stop word”, so our modele ignores the common words and focuses more on words that have a more significant impact on the prediction. Unfortunatly I was not able to implement a “stop words”.**

Sources

Matplotlib bars:
https://www.w3schools.com/python/matplotlib_bars.asp

Count plot:
https://seaborn.pydata.org/generated/seaborn.countplot.html

Insperation of binary classfication and Spam Filter:
https://www.youtube.com/watch?v=ij13PSnNdzw&t=902s