**Senti****ment Analysis Using Machine Learning**

# Data Preprocessing

Data preprocessing is a critical step in any machine learning pipeline. It involves transforming raw data into a format suitable for analysis.

## Importing Libraries

We begin by importing essential libraries like `pandas` for data manipulation, `nltk` for text processing, and `sklearn` for model building.



## Stop Words Remove

I download a file in which all English stop words from GitHub by using this link <https://github.com/stopwords-iso/stopwords-en/blob/master/stopwords-en.txt>  
  

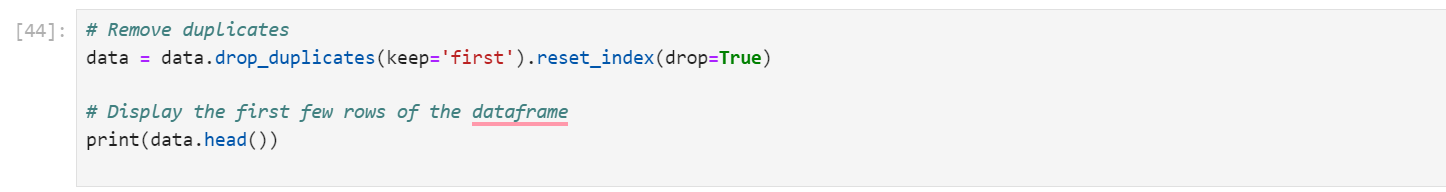

## Loading and Cleaning the Data

The IMDB dataset, containing movie reviews and their corresponding sentiment labels, is loaded. The text data is then cleaned by converting to lowercase, removing punctuation, special characters, and numbers.



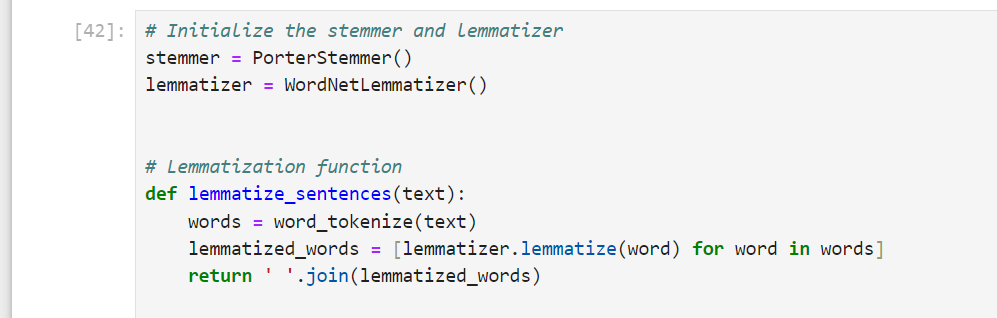
## Removing Duplicates

Duplicate reviews are removed to ensure that the model is trained on unique data.



## Tokenization and Lemmatization

The reviews are tokenized into individual words, and each word is lemmatized to reduce it to its base form. This step helps in normalizing the text data, making it more suitable for machine learning models.

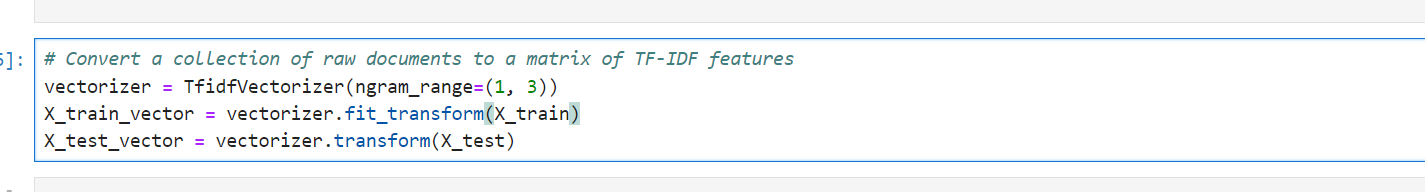


# Feature Extraction

Feature extraction involves transforming the text data into numerical features that can be used to train a machine learning model.

## TF-IDF Vectorization

The text data is transformed into a TF-IDF (Term Frequency-Inverse Document Frequency) matrix. This matrix represents the importance of words in the corpus, with higher values indicating more significant terms for classification.

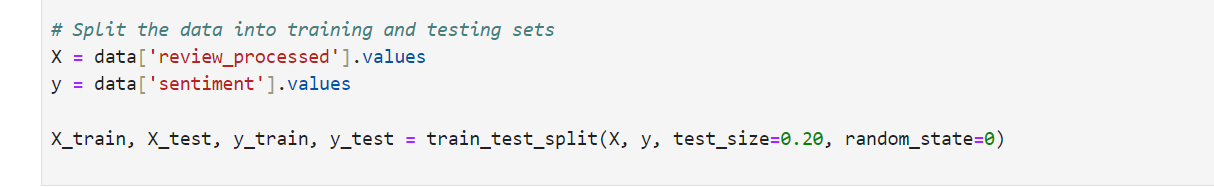


# Model Training and Evaluation

This section covers the training of machine learning models, making predictions, and evaluating model performance.

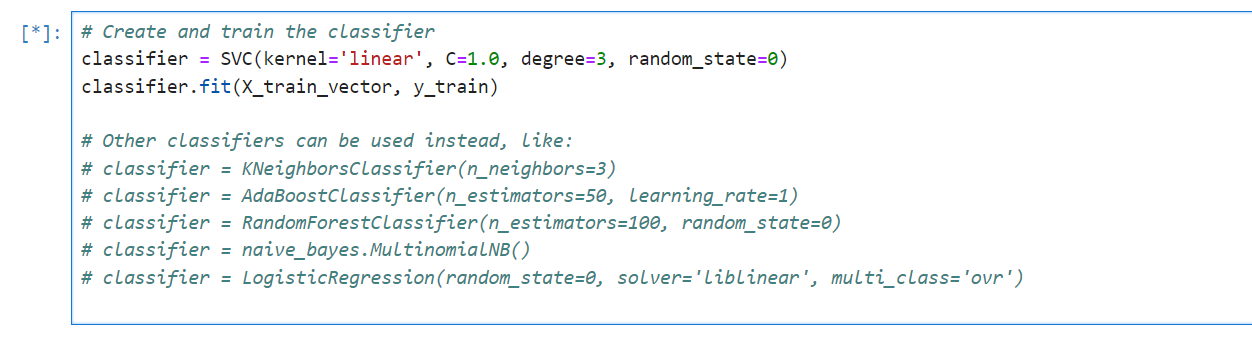
## Splitting the Data

The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.



## Training the Classifier

Several classifiers, such as Support Vector Machines (SVM), K-Nearest Neighbors, and Random Forest, are trained on the TF-IDF features. The report focuses on the SVM model with a linear kernel.  
I utilized several machines learning models, including KNN, Naïve Bayes, Logistic Regression, and AdaBoost. To evaluate each model, simply uncomment the desired model, comment out the others, and run the code.



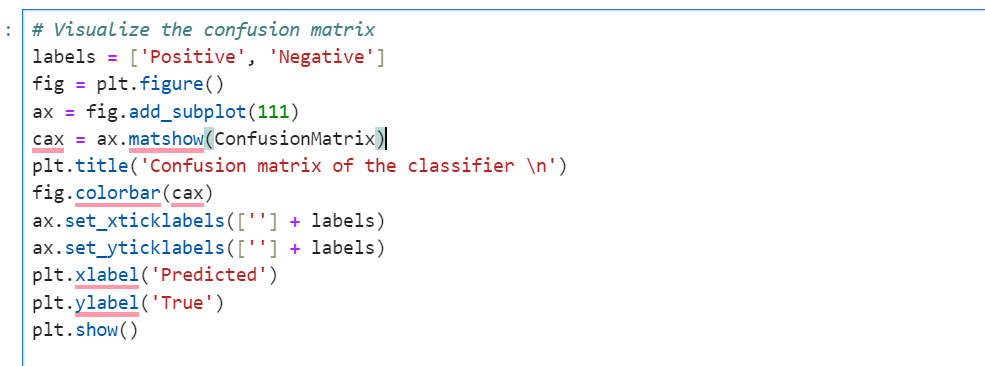
## Prediction and Evaluation

The trained model is used to predict sentiments on the test data. The predictions are then compared with the actual labels to evaluate the model's performance.



## Confusion Matrix and Classification Report

The confusion matrix is visualized, and the classification report, which includes precision, recall, and F1-score, is generated to assess the model's effectiveness.



# Conclusion

The conclusion summarizes the findings from the sentiment analysis, discussing the accuracy achieved and areas for potential improvement. Future work could involve experimenting with more advanced models or different preprocessing techniques.