Sentimental Analysis on Movie Reviews

Sentiment Analysis, also known as Opinion Mining, involves classifying data (text) into positive, negative, or neutral sentiments. Below are the steps involved in building a Sentiment Analysis model for movie reviews using the IMDb dataset:

1. Data Collection

The imdb dataset movie review dataset is collected from Kaggle. The data have 50,000 rows with column names of 'review' and 'sentiments'. The target column is the Sentiment with 'Positive' and 'Negative' sentiments. The classification algorithm is used to predict the proper class of the sentiments. This dataset will be used for training and testing the model.

2. EDA (Explanatory Data Analysis)

It is used to understand the characteristics of data. It involves loading and inspecting data as a data frame. The data is loaded using 'Pandas' library. Also, necessary libraries are installed and loaded on Google Colab. It allows to get the dataframe information, categories of target columns, null value checking, find length of each review etc. The goal of EDA is to gain insights from data, identify patterns and then to take pre-processing decisions.

3. Data Preprocessing

It involves cleaning the dataset. The input data is the 'Review' data. Cleaning of the text is done by removing URLS, HTML tags, contractions of words, special characters, stop words (to reduce noise in the data), emojis etc, most frequent words. This should be removed as it can affect the overall accuracy of the model. Using the code, the count and words are predicted in order to do preprocessing.

After cleaning the model, the text should be lowered to ensure data consistency, lemmatized the word to reduce the words to their base form and tokenize the text to break the review into individual words. After the processing of above text, the 'Target' column should be converted to numerical form. The computer only understands the numerical data. There are 2 sentiments. For negative sentiments, it is classified as '0' and for positive sentiment, its converted to '1'. Then mapping gets finished. After the conversion to numerical/ Label encoding form, the review field needs to be tokenized for further processing for splitting the dataset into training and testing data.

4. Feature Extraction

There are 2 techniques like TF-IDF (Term Frequency- Inverse Document Frequency) and CountVectorization to convert the text review data into numerical features.

Using TF-IDF vectorization, the accuracy of the model was reduced to a great extent. So, choose the CountVectorizer technique. CountVectorizer is a simple method that counts the occurrences of each word in a document. It removes common English stop words (e.g., 'the', 'and' 'is'). It considers only unigrams (single words) during the tokenization process. It applies the CountVectorizer to the input text data and transforms it into a matrix of token counts. using fit_transorm method. Then the data is split it into training and testing data. For training data, there is 80% data and for testing, it used 20% of overall data.

5. Model Building

Since the problem is a classification problem, the classification algorithm commonly used are Naive Bayes classification algorithm, Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier. The models are fitted using, fit method to fit the training and testing reviews. The model also predicts the label to classify it to Positive or Negative sentiments.

6. Model Evaluation

Using the testing data, the model is evaluated to calculate the and make relevant metrics to calculate accuracy, precision, recall, F1-score. Classification report have been created to predict the metrics. The confusion matrix also depicted to show the True Positive and True Negative cases.

For Naive Bayes model, accuracy of the model is 87%.

For Logistic Regression model, accuracy of the model is 88%.

For Random Forest Classifier, the accuracy of the model is 86%. We can do Hyper-parameter tuning to get more accuracy.

Since all the models can be considered as a 'Perfect' model. The most accurate model among the 3 is the Logistic Regression of accuarcy 88%.

Inference: The data have been trained using 50000. The data is very small to get better accuarcy. More data, more accurate the data will be.

7. Predicting the unseen data

new_review = "This movie was great! I loved the storyline."

Steps to predict the sentiment of unseen data:
1. Preprocess the data (def cleaning ())
2. Removal of stop words, frequent words.
3. Perform lemmatization, tokenization, Count-vectorization using transform and cv taken.
4. Application of logistic regression as it is accurate.

5. Convert numerical predictions to text labels using Label Encoder/ Binarize or mapping to numbers.

Result: Predicted Sentiment- positive.