מכללה אקדמית כנרת

מרצה: פרופ' מאלכ יוסף

הערות:

ההגשה היא בזוגות
מאוד חשוב לפתור את הפרויקט באופן עצמאי
תהיה הגנה על הפרויקט בפני המרצה
יש להגיש מחברת עם תיעוד מלא
קובץ הדאטא צריך להיות באותה תיקייה של מחברת הפתרון- יש להגיש שניהם ביחד כתיקייה דחוסה

Text Classification for Customer Reviews

Data :

https://www.kaggle.com/datasets/vigneshwarsofficial/reviews

Problem Description:

You are working for a company that sells electronic products online. The company receives a large number of customer reviews for its products. For example for **Restaurant Customer Reviews**. Your task is to develop a text classification model that can automatically classify customer reviews into different categories based on their sentiments (e.g., positive, negative, neutral).

Dataset:

The company has provided you with a dataset containing customer reviews along with their corresponding labels (sentiments). Each review is a text document, and each label represents the sentiment associated with the review.

Part 1: Data Loading and Preprocessing

1. **Load the Data:**

- Obtain the customer review dataset from a reliable source (e.g., UCI Machine Learning Repository, Kaggle, Amazon Customer Reviews Dataset).

2. Data Cleaning and Preprocessing:

- Perform necessary cleaning steps, such as removing irrelevant information (e.g., HTML tags, special characters) from the text data.

- Convert all text to lowercase to ensure consistency.

- Apply lemmatization to reduce words to their base form.

- Remove stop words from the text data to eliminate commonly used words that may not contribute much to sentiment classification.

>> You might save the output to a nee file-This new file can be use for the second part.

>> Or you can merge the two parts

Part 2: Text Classification with Pipeline

3. Define a Pipeline:

- Create a Scikit-learn pipeline that incorporates the necessary preprocessing steps and text classification algorithm.

4. Split the Data:

- Split the preprocessed data into training and testing sets for model development and evaluation.

5. Model Training and Evaluation:

- Configure the pipeline to train three different text classification algorithms on the training data (e.g., Naive Bayes, Support Vector Machines, Random Forest).
- Evaluate the performance of each trained model using appropriate evaluation metrics (e.g., accuracy, precision, recall) on the testing data.
- Summarize all the results into suitable table.

This is an example of such summarize table:

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.85	0.87	0.82	0.84
Random Forest	0.89	0.91	0.87	0.89
Support Vector Machines	0.91	0.89	0.92	0.90
Logistic Regression	0.88	0.90	0.86	0.88
Gradient Boosting	0.92	0.93	0.91	0.92

6 A report or Jupyter notebook documenting the entire project, explaining the steps taken, discussing the findings, and providing insights into the challenges faced.