NLP MINI PROJECT

Name : Murali Kumar R

Roll No : 225229120

Preprocessing Techniques :

- ✓ Check Missing values
- ✓ Drop Unwanted Column
- ✓ Converting text to lowercase
- ✓ Removing stopwords from the data
- ✓ Removing punctuation marks
- ✓ Removing special characters

Algorithms :

Random Forest Classification:

- ✓ Random Forest Classification can also be used in Natural Language Processing (NLP) tasks such as text classification, sentiment analysis, and named entity recognition.
- ✓ These vectorized representations can be used as input to a Random Forest Classification model to predict the class labels of the text.
- ✓ We evaluated the model's performance on the test set using several metrics, including accuracy. The model achieved an accuracy of **0.085.**

Support Vector Classification :

- ✓ SVC is a type of supervised learning algorithm that works by finding a hyperplane that separates the data into different classes.
- ✓ In NLP, text data is often represented using vectorized representations such as Bag-of-Words, TF-IDF, or Word Embeddings. These vectorized representations can be used as input to an SVC model to predict the class labels of the text.
- ✓ To train an SVC model for text classification, we first need to vectorize the text data. This is done using various techniques such as Bag-of-Words or TF-IDF.

Once the data is vectorized, we can split it into training and test sets and train the SVC model on the training data.

- ✓ We evaluated the model's performance on the test set using several metrics, including accuracy. The model achieved an accuracy of **0.09.**

Decision Tree Classification :

- ✓ A decision tree is a flowchart-like structure where each internal node represents a test on an attribute or feature, each branch represents the outcome of the test, and each leaf node represents a class label.
- ✓ In NLP, text data is often represented using vectorized representations such as Bag-of-Words, TF-IDF, or Word Embeddings. These vectorized representations can be used as input to a decision tree model to predict the class labels of the text.
- ✓ Additionally, ensemble methods such as Random Forests and Boosted Trees can be used to improve the performance of decision tree models.
- ✓ We evaluated the model's performance on the test set using several metrics, including accuracy. The model achieved an accuracy of **0.075.**

Evaluation metrics and Model tuning:

Accuracy score :

In the example provided, we are comparing the accuracy scores of three different machine learning models (Random Forest, Support Vector Machine (SVM), and Decision Tree) on a particular NLP task :

- ✓ Random Forest Classification : **0.085**
- ✓ Support Vector Classification : **0.09**
- ✓ Decision Tree Classification : **0.075**

Therefore, the best model is : Support **Vector Classification.**