Text Classification

As a data scientist, you've been tasked with developing a text classification model to identify whether a given tweet is about a real disaster or not. You will work with the "Disaster Tweets" dataset, which contains a collection of tweets labeled as either related to real disasters or not.

Your objective is to build and evaluate multiple text classification models using different feature engineering techniques to accurately classify tweets into two categories: "real disaster" and "not real disaster". This model will aid in automatically flagging tweets that report actual emergencies or disasters, helping emergency response teams and authorities to quickly identify and respond to critical situations.

Consider the following columns:

text: The content of the tweet.

target: Binary label indicating whether the tweet is about a real disaster (1) or not (0).

Tasks:

1. Data Preprocessing
2. Exploratory Data Analysis (EDA): Conduct exploratory data analysis to understand the distribution of the target classes and the characteristics of the text data. Visualize the distribution of target classes and explore the most common words or phrases in both categories.
3. Feature Engineering: Experiment with different feature engineering techniques such as:

   Bag-of-Words (BoW)

   TF-IDF (Term Frequency-Inverse Document Frequency)

   Word Embeddings: Use pre-trained word embeddings (e.g., Word2Vec, GloVe)

   Character-level Features: Extract features based on character n-grams to capture patterns in the text.

4. Model Selection: Train and evaluate multiple classification models using the engineered features. Experiment with models such as Logistic Regression, Naive Bayes, Decision Trees, Random Forests
5. Model Evaluation: Evaluate the performance of each model using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Compare the performance of different models and feature engineering techniques to identify the most effective approach.
6. Hyperparameter Tuning: Fine-tune the hyperparameters of the best-performing models using techniques like grid search or random search to optimize their performance further.
7. Model Interpretation: Interpret the results of the best-performing model to understand which features contribute most to the classification task.