NLP

Disaster Tweet Classification

1. Datasets: train.csv and test.csv

train shape : (5080,8)

test Shape : (3263,4)

Missing values:

some null values in the 'keyword' and 'location' columns

Target distribution:

- 0 (Not disaster-related): 2884

- 1 (Disaster-related): 2196

1. Data Preprocessing

Data Cleaning and Feature Engineering

Null values drop from the training set

Text preprocessing function `clean\_text()` implemented:

* Removed HTML tags and URLs

Removed special characters

Converted text to lowercase

Tokenization

Removed stopwords

Stemming and lemmatization

Additional features created:

tweet\_length: Length of the tweet

num\_words: Number of words in the tweet

num\_sentences: Number of sentences in the tweet

3) Feature Extraction

TF-IDF Vectorization used to convert text data into numerical features

4) Model Creating and Evaluation

Four different models train and evaluate:

**-Support Vector Machine (SVM)**

Accuracy: 0.80

F1-score (weighted avg): 0.79

Precision: 0.80

Recall: 0.80

**- Logistic Regression**

Accuracy: 0.79

F1-score (weighted avg): 0.78

Precision: 0.79

Recall: 0.79

**- Random Forest**

Accuracy: 0.77

F1-score (weighted avg): 0.76

Precision: 0.77

Recall: 0.77

**- Naive Bayes**

Accuracy: 0.60

F1-score (weighted avg): 0.60

Precision: 0.64

Recall: 0.60

The best Model is svm based on evaluation metrics based on evaluation metrics