**Project:** **Natural language processing with disaster tweets**

**Milestone-2**

Introduction:   
Natural Language Processing, also known as NLP, is a subfield of computer science, specifically artificial intelligence, that focuses on understanding written and spoken text. It covers various tasks some of which are speech recognition, sentiment analysis and language generation; And, it has been applied in several use cases such as machine translation, spam detection, virtual assistants and chatbots.

Datasets:

The primary dataset used in the **Natural language processing with disaster tweets**

Available on Kaggle, which includes information like id, keyword, location, target, text etc.

Literature survey related to the topic:

1. Title: Natural Language Processing with Disaster Tweets

Techniques/Tools: Python, scikit-learn, TensorFlow, Keras.

Preprocessing: Data cleaning and feature extraction.

Text vectorization using TF-IDF. Tokenization and padding of text sequences.

Models Used: Classical Machine Learning: Logistic Regression, SVM.

Deep Learning: LSTM (Long Short-Term Memory).

Performance: Classification of tweets as disaster or non-disaster.

Achieved a validation accuracy of 0.85 using LSTM.

Reference: [GitHub - zeyongj/Natural-Language-Processing-with-Disaster-Tweets: Kaggle Project: Predict which Tweets are about real disasters and which ones are not.](https://github.com/zeyongj/Natural-Language-Processing-with-Disaster-Tweets)

2. Title: Introduction to NLP with Disaster Tweets

Techniques/Tools: Python, NLTK, pandas, scikit-learn.

Preprocessing: Text cleaning (removal of URLs, special characters).

Tokenization and lemmatization.Vectorization using TF-IDF and word embeddings.

Models Used: Naïve Bayes, Logistic Regression.

Performance: Achieved a strong classification performance for tweet categorization.

Reference: [Introduction to NLP with Disaster Tweets | by Kheirie Elhariri | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/introduction-to-nlp-with-disaster-tweets-3b672a75748c)

3. Title: Detecting Disaster from Tweets: Classical ML and LSTM Approach

Techniques/Tools: Python, scikit-learn, Keras.

Preprocessing: Data cleaning, including URL removal and lowercasing.

Feature extraction using TF-IDF and word embeddings.

Models Used: Classical ML models: Logistic Regression, SVM.

Deep Learning: LSTM and Bidirectional LSTM.

Performance: Deep learning approach with LSTM outperformed classical ML models, reaching 0.92 accuracy.

Reference: <https://towardsdatascience.com/detecting-disaster-from-tweets-classical-ml-and-lstm-approach-4566871af5f7>

4. Title: Detecting Disaster Tweets using a Natural Language Processing Technique

Techniques/Tools: Python, TensorFlow, scikit-learn.

Preprocessing: Text cleaning and tokenization.

Removal of stopwords and non-alphabetic tokens.TF-IDF vectorization.

Models Used: LSTM, Word2Vec embeddings.

Performance: Achieved high performance in tweet classification with an LSTM model.

Best results obtained with LSTM using pre-trained word embeddings.

Reference:<https://www.researchgate.net/publication/356647529_Detecting_Disaster_Tweets_using_a_Natural_Language_Processing_technique>

5. Title: Sentiment Analysis: Predicting Whether a Tweet is About a Disaster

Techniques/Tools: Python, pandas, NLTK, scikit-learn.

Preprocessing: Data cleaning, tokenization, and lemmatization.Use of word embeddings for vectorization.

Models Used: Logistic Regression, Naïve Bayes, Support Vector Machine (SVM).

Performance: Prediction model built with 0.85 validation accuracy.Successful differentiation of tweets related to disasters from non-disasters.

Reference: <https://towardsdatascience.com/sentiment-analysis-predicting-whether-a-tweet-is-about-a-disaster-c004d09d7245>

Data Pre-processing:

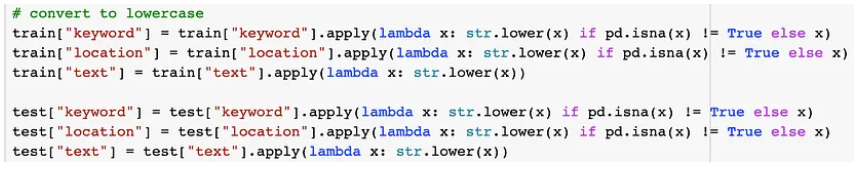
Handling missing values (e.g. Filling in missing values with mean)

> **Text Cleaning:**

Remove URLs, mentions, hashtags, and special characters.

Convert text to lowercase for uniformity.

>The “target” column served as our guiding star, with a value of ‘1’ indicating a disaster-related tweet and ‘0’ for non-disaster tweets.

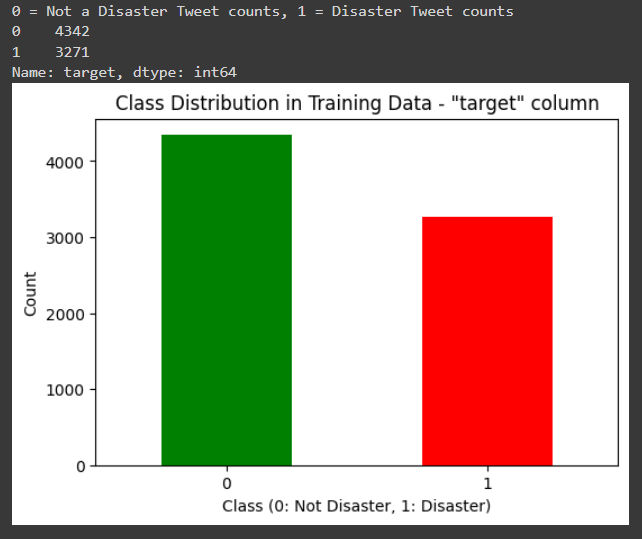


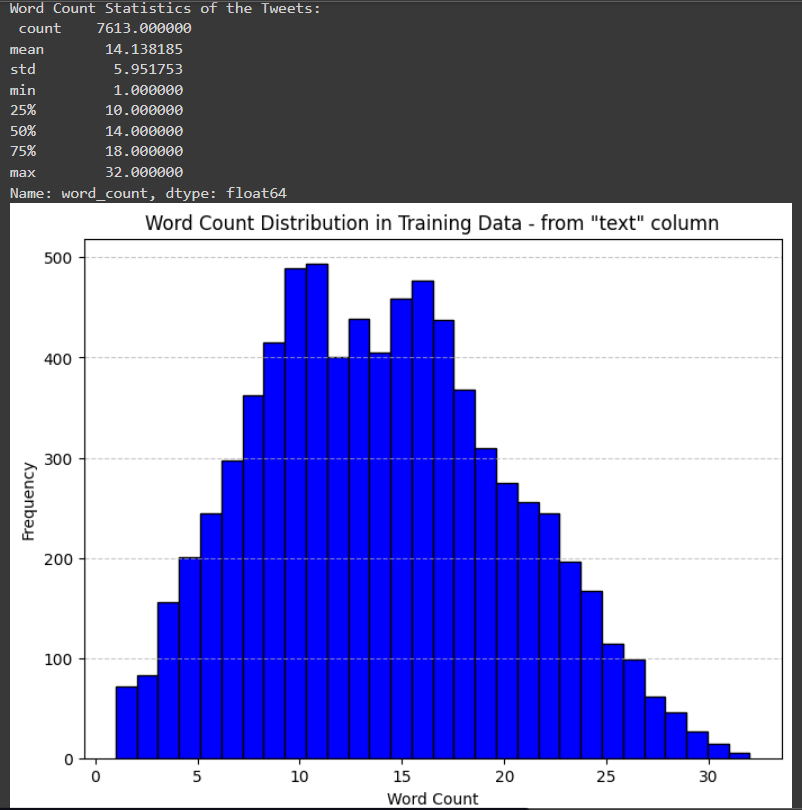


Data Visualization / EDA Python Lib:

-Matplotlib is used for displaying detection in the notebook, helping visualize each detection step-by-step.

- I meticulously calculated word counts in our tweets and identified an optimal maximum sequence length to ensure our text didn’t get cut off or padded unnecessarily during encoding.





Model Creation and Testing :

->Using logistic regression to predict the binary outcome of survival

->Training the model on the training data

->After preprocessing the data and splitting it into X\_ train, X\_ test, y\_ train, and y\_ test, the assistant creates a logistic regression model using scikit-learn

The figure below shows the result of the model fitting stopping at epoch 23 with a training accuracy of 0.8456 and validation accuracy of 0.8031.

