**Disaster Tweet Analyzer**

Progress from Week 1 to Week 2

**About Infosys Springboard Internship:**

This internship has been a meaningful experience for me. The daily sessions have helped me stay consistent with learning, and the structured curriculum ensured steady progress. Grateful to the mentor, who is always approachable for clearing doubts and assigning practical tasks. His guidance made it easier to connect theoretical concepts with real-world scenarios.

One of the highlights for me was the opportunity to collaborate with other learners. The knowledge-sharing environment fostered teamwork and gave me different perspectives, which enriched my learning. Overall, this internship has been a great introduction to professional life, helping me grow not just technically but also in collaboration skills.

**Submitted to:** Nitig Singh Sir

**Submitted by:** Deela Neha Reddy Thummuru

**Introduction**

**Project Overview:**

During emergencies and natural disasters, social media platforms become crucial channels for real-time information exchange and crisis communication. This project focuses on leveraging natural language processing (NLP) techniques to analyse tweets related to disasters and emergencies. By classifying tweets as either relevant to a disaster or not, the project aims to assist emergency responders and organizations in identifying critical information amidst the noise of social media.

**Objectives:**

1. Develop a text classification system capable of identifying tweets relevant to disasters or emergencies.
2. Explore and implement state-of-the-art NLP techniques for tweet preprocessing and feature extraction.
3. Evaluate the performance of the classification models across different disaster scenarios and datasets.
4. Showcase the practical applications of NLP in crisis communication and emergency response.
5. Document the project methodology, findings, and recommendations for knowledge sharing and future research.

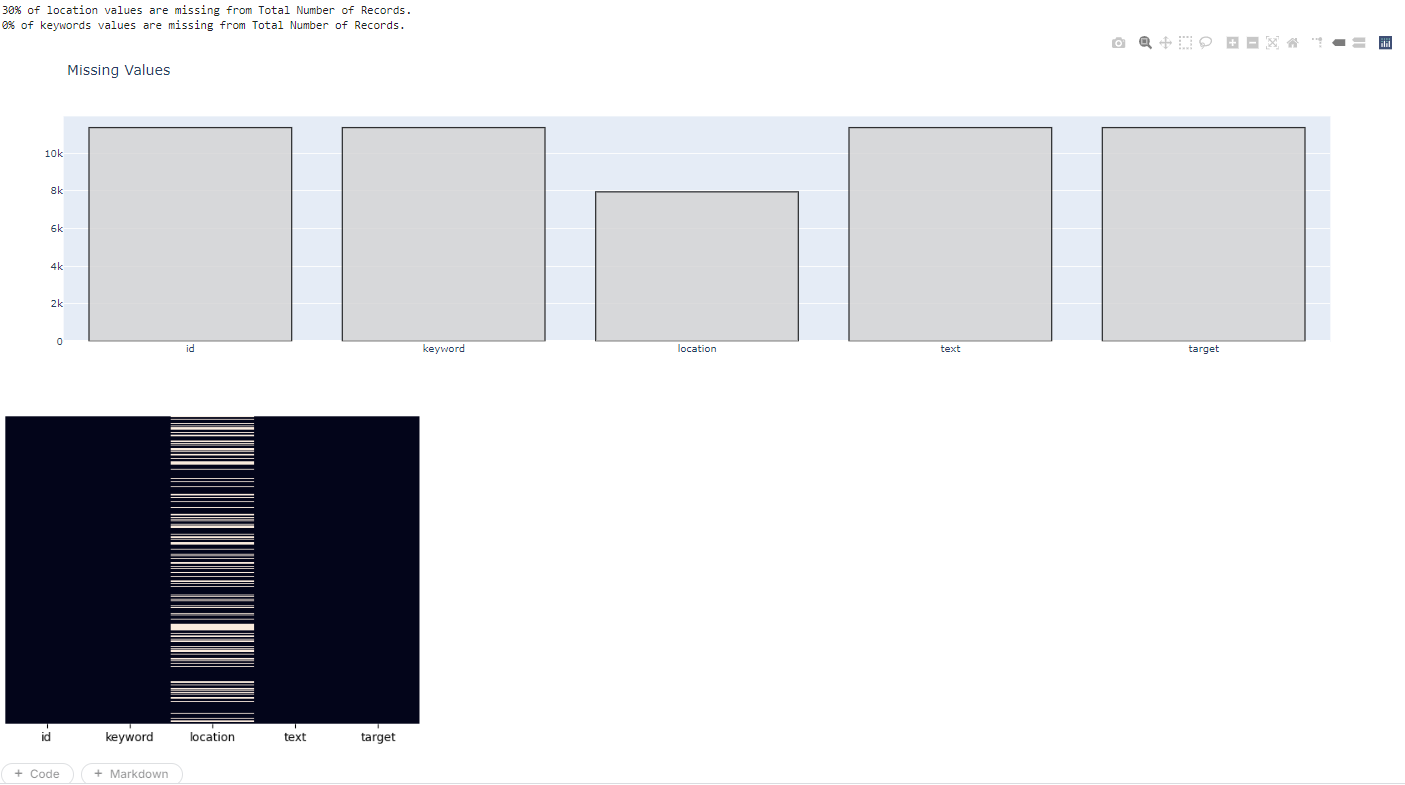
**Dataset and Methodology:**

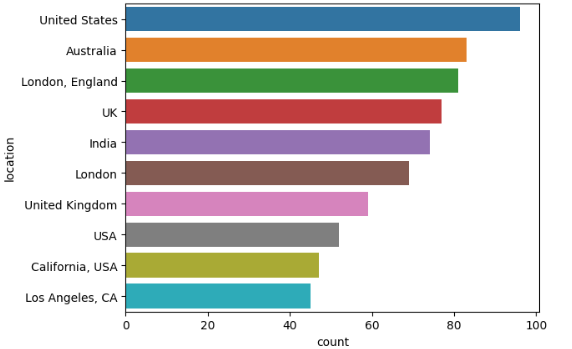
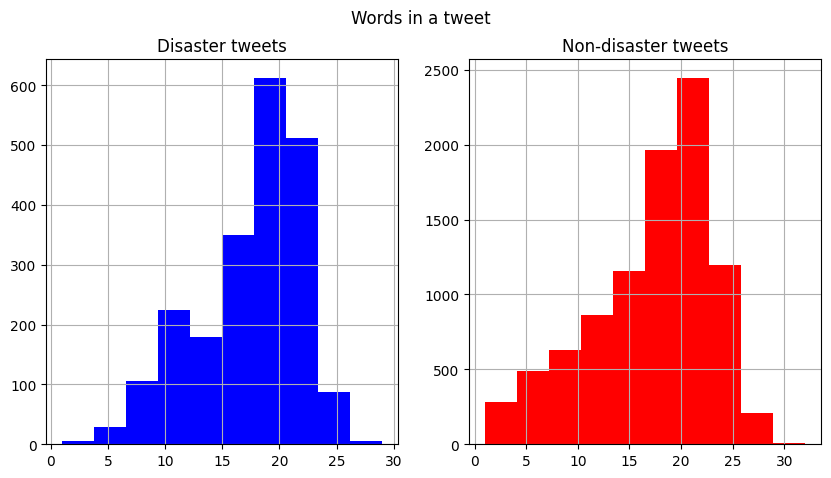
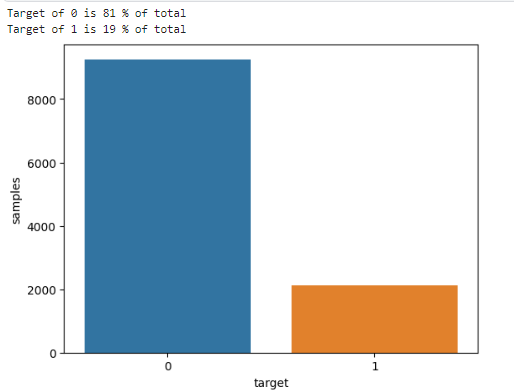
**Dataset:**The dataset used for this project was sourced from Kaggle and consists of tweets labelled as either disaster-related or non-disaster-related. This dataset serves as a solid foundation for performing classification of tasks.

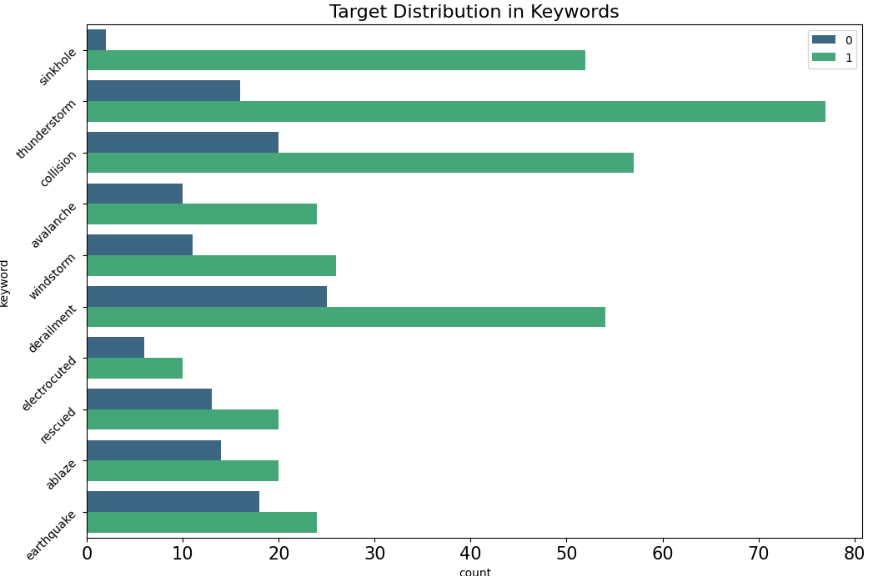
**Methodology:**

* **Data Exploration:**
  1. Displayed the first five rows of the dataset to understand its structure and content.
  2. Showed basic statistics of the dataset, including mean, standard deviation, minimum, and maximum values.
  3. Provided information about the dataset, such as data types and the number of non-null values for each column.
  4. Assessed the shape of the dataset to determine the number of rows and columns.
  5. Checked for any missing values in the dataset to identify potential data integrity issues**.**
* **Data Cleaning:**
  1. Converted all text to lowercase to maintain uniformity across the dataset. This ensures that variations of the same word are treated equally during analysis.
  2. Removed stop words to eliminate common terms that do not contribute significant meaning to the text, allowing the model to focus on more relevant words.
  3. Eliminated emojis to simplify the dataset and avoid ambiguity in sentiment classification.
  4. Removed punctuation marks to standardize the text and prevent the model from misinterpreting punctuation as meaningful data.
  5. Cleared special characters to reduce noise and ensure that only relevant words are processed.
  6. Removed URLs from the tweets to maintain focus on the textual content.
  7. Dropped the 'location' column from the training dataset as it was deemed unnecessary for the analysis.
* **Visualization:**
  1. Created diagrams to visually represent missing values, providing insights into data completeness and highlighting any potential issues.
  2. Visualized the distribution of word counts in comparison to the target feature (disaster vs. non-disaster tweets) to understand how word usage varies across classes.
  3. Analysed the distribution of text lengths relative to the target feature, offering insights into how tweet length might affect classification outcomes.
  4. Represented the most frequently occurring keywords, highlighting significant terms in both disaster and non-disaster contexts.
  5. Examined the target distribution concerning keywords, revealing how specific words are associated with disaster-related versus non-disaster tweets.
  6. Represented the total number of disaster and non-disaster tweets available in the dataset, identifying any class imbalance.
  7. Analyzed the distribution of tweets based on their location data to understand potential regional trends in disaster reporting.
  8. Compared specific words used in disaster and non-disaster tweets, identifying distinctive linguistic patterns.
  9. Visualized the distribution of tweet lengths specifically for disaster and non-disaster tweets, allowing for a focused analysis of how tweet length influences classification.

**Results:**

****

* 



**Conclusion:**

The Disaster Tweet Analyzer project effectively explored a dataset of tweets to classify them as disaster-related or non-disaster-related. Through comprehensive data exploration, key tasks such as displaying initial rows, assessing data integrity, and analysing basic statistics were conducted. The dataset was meticulously cleaned by converting text to lowercase, removing stop words, emojis, punctuation marks, special characters, and URLs. Visualizations provided critical insights, including the distribution of word counts, text lengths, and the most repeated keywords, which revealed linguistic patterns indicative of the tweet classes. This groundwork sets the stage for future analysis and machine learning applications aimed at enhancing crisis communication and response strategies.

**Future Objectives for the Next Two Weeks:**

**Model Development:** Begin developing machine learning models for classifying tweets. You might explore various algorithms, such as Logistic Regression, Random Forest, or more advanced techniques like Support Vector Machines (SVM) and neural networks. Evaluate the models using metrics such as accuracy, precision, recall, and F1 score.

**Hyperparameter Tuning:** Optimize the performance of your selected models through hyperparameter tuning. Techniques like Grid Search or Random Search can help identify the best parameters for your models.

**References:**

**IPYNB file**: [Kaggle Dataset](https://www.kaggle.com/code/deelanehareddy/notebookacc1533072/edit)