Disaster Tweets Classification Kaggle NLP Challenge

Overview

This document outlines our solution for the **"Real or Not? NLP with Disaster Tweets"** Kaggle competition. The challenge involves **Natural Language Processing (NLP)** to classify tweets as either:

- **1 (Real Disaster)** Tweet indicates an actual disaster event.
- **0 (Not Disaster)** Tweet is unrelated or uses disaster terms metaphorically.

Twitter is a critical platform for emergency communication. This project aims to assist **disaster relief organizations** and **news agencies** in filtering relevant tweets for faster response.

Team Members

- Dhairya A Mehra
- Jay Gondaliya
- Pranay Vasoya
- Praneet Mahendrakar

Dataset

The dataset, provided by Kaggle, contains **10,000 labeled tweets** and is split into:

- **train.csv** Labeled data for training the model.
- **test.csv** Unlabeled data for predictions.
- **sample_submission.csv** Format for Kaggle submission.

Dataset Link: https://www.kaggle.com/competitions/nlp-getting-started

Problem Statement

The objective is to build a **binary text classification model** to determine whether a tweet refers to a real disaster. For example:

Tweet Text	Target
Forest fire near La Ronge Sask. Canada	1
My phone battery is on fire	0

Tech Stack

- **Language:** Python 3.x
- **Libraries:**
  - `pandas`, `numpy` Data handling
  - `matplotlib`, `seaborn` EDA & visualization
  - `nltk`, `re` Text preprocessing
  - `scikit-learn` Feature extraction & model building
  - `tensorflow` / `pytorch` (optional) Deep learning approaches
- **Platform:** Kaggle Notebooks

Approach

1. Data Preprocessing

- Remove URLs, mentions, hashtags, and special characters.
- Convert text to lowercase.
- Tokenization, stopword removal, and stemming/lemmatization.

2. Feature Engineering

- **TF-IDF Vectorization**
- **Word Embeddings** (GloVe, Word2Vec)

3. Modeling

- **Baseline:** Logistic Regression, Naive Bayes
- **Advanced:** LSTM, BERT-based models

4. Evaluation

- **Metric:** F1-score (primary for imbalanced data)
- Cross-validation to prevent overfitting

5. Submission

- Generate `submission.csv` in Kaggle format

Performance Metrics

Model	F1 Score
Logistic Regression	TBD
LSTM	TBD
BERT	TBD

How to Run

```
# Clone the repository
git clone https://github.com/<your-repo>.git
cd disaster-tweets-classification

# Install dependencies
pip install -r requirements.txt

# Run the notebook
jupyter notebook Disaster_Tweets_NLP.ipynb
```