Natural Language Processing Foundation

Project-1

Tasneem mohamed mohamed mohamed

1. Project Overview

The goal of this project was to classify tweets as either related to natural disasters or not. We used Natural Language Processing (NLP) techniques and machine learning models to achieve this, aiming to support disaster response by quickly identifying relevant social media posts.

---

2. Dataset

- Source: Kaggle NLP Getting Started dataset
- Size: 7,613 tweets with 5 columns (id, keyword, location, text, target)
- Target distribution: 4,342 disaster-related tweets (target=1) and 3,271 non-disaster tweets (target=0)

---

3. Data Preprocessing

- Removed noise: special characters, URLs, and HTML tags
- Converted text to lowercase
- Removed stopwords using NLTK
- Applied **lemmatization**
- Tokenized text into individual words
- For feature extraction, used **TF-IDF** and additional features like tweet length and presence of disaster keywords

---

4. Model Training

- **Models used**:
    1. Naive Bayes
    2. Logistic Regression
    3. Support Vector Machine (SVM)
    4. DistilBERT (Transformer-based model for NLP)
- Split data into train and validation sets (80/20 split)
- Used cross-validation and hyperparameter tuning for classical ML models

---

5. Model Evaluation

- Classical ML models reached **~80–81% accuracy** using TF-IDF features
- Adding extra features improved accuracy slightly to **81%**
- **Word embeddings (GloVe)** produced lower accuracy (~75%) on the same split
- **DistilBERT** achieved the best performance on train/validation set:
    - **Validation Accuracy:** 83%
    - **Precision/Recall/F1:** Disaster class ≈ 82%/77%/80%
    - **Confusion Matrix:** correctly classified most examples, some misclassifications remain

Note: Accuracy on the full training set was higher (~95%) due to overfitting, but validation set provides a more realistic measure of generalization.

---

6. Dashboard

A dashboard was created inside the notebook using **matplotlib, seaborn, and pandas**:

- **Heatmap** for the confusion matrix
- **Classification Report** as a styled DataFrame
- **Bar chart** showing distribution of disaster vs non-disaster tweets
- **Interactive input box** to test DistilBERT on any new tweet

This allows easy visual inspection of model performance and experimentation on new examples.

---

7. Conclusion

- Preprocessing and feature engineering are crucial for ML models
- DistilBERT significantly outperforms classical ML models on text classification
- Validation metrics provide a realistic measure of generalization
- Dashboard provides a professional and interactive way to showcase results