NLP PROJECT FOR DISASTER TWEET CLASSIFICATION

PRESENTED BY: DEEPAK SHARMA

INTRODUCTION

- Objective: Develop a machine learning model to classify tweets into disaster and non-disaster categories using natural language processing techniques.

DATA OVERVIEW

- The dataset contains 7,613 tweets with columns for id, keyword, location, text, and target. Key challenges include missing values and the need for text normalization.

DATA CLEANING

- Steps included:
- - Removing URLs, HTML tags, and special characters.
- - Converting texts to lowercase.
- - Handling missing values in 'keyword' and 'location'.

FEATURE ENGINEERING

- Features extracted:
- - Word counts and TF-IDF scores using CountVectorizer and TfidfVectorizer.
- - Sentiment polarity scores.
- - Additional features like tweet length, hashtags, and mentions.

EXPLORATORY DATA ANALYSIS

- Univariate analysis showed balanced classes and insights into tweet lengths. Sentiment analysis revealed differences in sentiment polarity between disaster and non-disaster tweets.

MODEL SELECTION

- Evaluated Logistic Regression and Random Forest classifiers. Initial testing with cross-validation showed Logistic Regression as more promising.

MODEL TRAINING

- Used TF-IDF features for training. Cross-validation scores helped compare model performance before tuning.

HYPERPARAMETER TUNING

- Applied Grid Search CV to optimize Logistic Regression. Simplified the Random Forest model to reduce complexity.

MODEL EVALUATION

- Final evaluation on test data:
- - Accuracy: 80.43%
- - Precision: 82.08%
- - Recall: 69.18%
- - F1-Score: 75.08%
- Confusion matrix analysis provided insights into model performance.

RESULTS AND INSIGHTS

- The logistic regression model effectively classified disaster tweets with high precision, demonstrating the importance of feature engineering and model tuning in NLP tasks.

CHALLENGES AND LEARNINGS

- Challenges included handling missing data and imbalanced classes. Learnings emphasized the value of comprehensive data cleaning and the impact of feature selection on model accuracy.

FUTURE WORK

- Future improvements could include exploring more complex NLP models like BERT, incorporating more granular sentiment analysis, and using larger, more diverse datasets to enhance model robustness.

CONCLUSION

- This project underscores the potential of machine learning in disaster response scenarios, highlighting how NLP can be leveraged to quickly classify critical information from social media.