

NLP Sentiment Analysis Project Summary

Introduction

This project focuses on sentiment analysis of disaster-related tweets using Natural Language Processing (NLP) techniques. It aims to classify tweets as either disaster-related or non-disaster-related, aiding emergency response efforts.

Dataset and Tools

Dataset: Tweets labeled with disaster-related or non-disaster-related categories.

Tools: Python libraries including pandas, nltk, scikit-learn, and TensorFlow.

Data Preprocessing

Data is cleaned, tokenized, and preprocessed to remove stopwords, special characters, and perform stemming.

Balancing of data classes is achieved using RandomUnderSampler from the imbalanced-learn library.

Feature Engineering

Text data is vectorized using TF-IDF vectorization to convert text into numerical features.

Both Logistic Regression and Support Vector Machine (SVM) models are trained and evaluated for classification.

Model Development

Logistic Regression: Simple yet effective for text classification tasks.

SVM: Different kernel functions such as linear, polynomial, and radial basis function (RBF) are experimented with.

Model Evaluation

Models are evaluated using various metrics including accuracy, precision, recall, and F1-score.

Classification reports and confusion matrices are generated to assess model performance.

Prediction

Trained models are used to predict labels for unlabeled test tweets.

Predicted labels are stored in submission files for further analysis or deployment.

Conclusion

This project demonstrates the application of NLP techniques for sentiment analysis of disaster-related tweets. By preprocessing data, engineering features, and training logistic regression and SVM models, accurate predictions can be made, aiding disaster response efforts.

Future Scope

Potential enhancements include exploring advanced NLP techniques, optimizing model hyperparameters, and deploying the model for real-time tweet classification.