Natural Language Processing with Disaster Tweets

Capstone Project Proposal

Project Background

Natural Language Processing with Disaster Tweets is a Kaggle Challenge where tweets are collected with labels indication whether the tweets are about a disaster that occurred or not. Since tweets are social media language, therefore, it is a challenge to automatically identify them. Besides, ambiguity in texts makes it more difficult to achieve automatic identification of tweets containing information on real disaster.

Problem Statement

Given a tweet, this task is designed to identify if it contains information on occurrence of a real disaster or not.

Dataset and Inputs

Dataset Link on Kaggle: https://www.kaggle.com/c/nlp-getting-started/data

This dataset contains:

1. Training Data with 7613 instances
2. Testing Data with 3263 instances

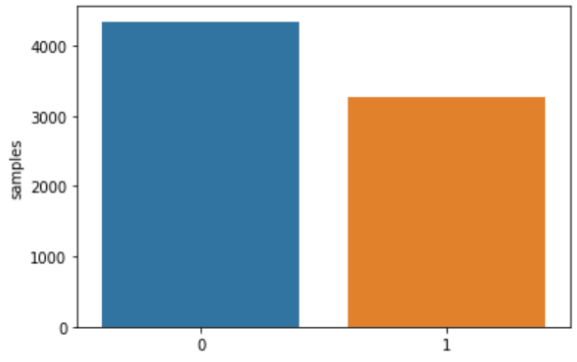
The train:test data instance ratio is 2:1.

This *train dataset* contains five columns:

- id – The unique identifier
- keyword – A keyword in that tweet
- location – The location from where the tweet was posted
- text – The tweet contents
- target – Indicates if the tweet is about a real disaster or not

The columns 'keyword' and 'location' contain missing values.

The test dataset contains four columns excluding the 'target' column from the train data.



In the 'target' column, there are two classes – 0, which indicates that the tweet is not related to a real disaster and 1, which indicates otherwise.

The data is not balanced where there is a difference of approximately 1000 tweets between the count of tweets about real disasters and the ones which are not. I plan to rather use stratified splitting than up or down sampling for this ratio of target labels which is about 4:3.

Solution Statement

This challenge is a classification problem. I will be using NLP techniques to design the classifier to identify if a tweet is about real disaster or not. An F1-score of more than 80% on the test set of this dataset is the target I would like to achieve from this project.

Benchmark Model

A logistic regression model with default configurations will be considered as the benchmark model. For improvements, I will be looking at other classification models and hyper-parameter tuning for improving predictability of the model.

Evaluation Metrics

F1 score will be primary metric that will be used to evaluate model. Since the training data is unbalanced, the accuracy will not be considered as a useful metric and only F1-score will be considered to avoid accuracy paradox. F1-score is helpful to monitor both precision and recall as metrics. An F1-score of more than 80% on the test data is the target I would like to achieve from this project.

Project Design

The nature of texts in conversational social media language which might contains special characters as well and might not be standard lexical English language. Therefore, pre-processing them is the key to designing a successful predictor. The different steps involved here are –

1. Exploratory Data Analysis
2. Data Pre-processing
    a. Removing HTML tags
    b. Extracting emoticons
    c. Spell Check
    d. Expanding contractions
    e. Stemming
3. Feature engineering using TF-IDF approach
4. Train benchmark model (Logistic Regression) and evaluate
5. Train other models to compare and tune performance
6. Final evaluation and comments