1. Problem Statement

This project seeks to answer the research question:

Can a pre-trained transformer-based model (i.e. a large language model or LLM) be used to improve the performance of non-transformer-based (NTB) classifier models (e.g. logistic regression or a neural network classifier) by augmenting its original training data?

To answer this question, the following hypothesis is used to guide the kind of data required from the analysis:

```
(performance of non-augmented training data) <
(performance of uninformed augmented training data) <
(performance of informed augmented training data)
```

Definitions

non-augmented - This is base case where only the originally provided data is used to train the NTB models

uninformed augmented - In this scenario, the original training data is doubled in size by prompting an LLM to provide a sample with similar sentiment to each of the original training samples.

informed augmented - In this scenario, the original training data is also doubled in size, but using a different prompt to the LLM

Note that the definitions of "*informed augmented*" may need to change based on what is discovered during the course of the project. If this definition is revised, it will be documented in the "Data profiling and quality report".

2. Data

Two types of data will be used in the project. The first type is referred to here as "base data" which is obtained from the ongoing kaggle competition titled "*Natural Language Processing with Disaster Tweets*" currently available at:

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

The second type of data is referred to here as "augmented data" which is generated from a LLM based on a particular prompt which is either uninformed or informed. The uniformed prompt will be designed to simply create a sample that is similar to each of the base data samples. The informed prompt will be designed to create similar samples but with additional conditions added such as using one of the top 10 most frequent words found in each class (disaster, not disaster).

This base data consists of two files: **train.csv** and **test.csv** which have the following characteristics:

train.csv - This file consists of 8562 lines of raw text. Each actual sample (row) has the following 5 fields (columns):

- **id -** integer, unique identifier for each row which should always have a value
- **keyword -** string, a particular keyword from the tweet which may be blank
- **location -** string, the location the tweet was sent from which may be blank
- **text -** string, the text of the tweet
- **target -** integer, 1 or 0 representing a binary label to be classified and denotes whether a tweet is about a real disaster (1) or not (0)

test.csv - This file consists of 3700 lines of raw text. Each actual test sample (row) has the same first 4 fields (columns) as the train.csv file: **id**, **keyword**, **location** and **text**. There is no target column because competition competitors are expected to predict and record this prediction as part of their submission.

Because test.csv data is unlabeled, it is only used at the end of the project to see how well each model performed under the experimental conditions described in the Methodology section.

Methodology

The first step in most NLP machine learning tasks is to pre-process the data into normalized tokens (lemmatized or stemmed words) and then encode these tokens

TODO

Resources

The project will be implemented in Python jupyter notebooks with a project module of helper functions intended to keep the notebook code clean and focused on the higher level aspects of the project. Free and open-source libraries such as sklearn and pytorch will be used to build the logistic regression and neural network models described in the Methodology section.

The following github project has been set up to do version control throughout development:

https://github.com/MichaelSzczepaniak/llmamd

A Hugging Face model is expected to be used as the LLM to generate the augmented data. Based on my current understanding of the terms of use, a free account should be sufficient to meet the goals of the project.