Machine Learning Engineer Nanodegree

Capstone Proposal

Srimaya Padhi
October 29th, 2020

Proposal

Domain Background

Natural language processing is a field where ML techniques are being aggressively pursued. Speech-to-text, text/social review monitoring, sentiment analysis, personal assistants are some of the applications of NLP. I got interested in NLP after coming across the deep learning driven text adventure “AIDungeon”. AIDungeon simulates role playing games by reacting to user issued inputs to create diverse storylines.

Problem Statement

The problem being solved is a classification problem of processing customer reviews from Trip Adviser and assigning it a numerical rating of 1 to 5.

Datasets and Inputs

Trip Adviser Hotel Reviews is the dataset being used in this project. The dataset was taken from Kaggle and is available under the Creative Commons license. The dataset consists user reviews for hotels mapped against a numerical rating of 1 to 5.

The dataset has 20491 reviews with an average rating of 3.95. The dataset has no null values. The ‘Review’ data is an English text-based field. It will be cleaned up and tokenized before being used as features. Generic words (connectors, prepositions, punctuation, etc.) will be removed to reduce the number of features and allow for faster training.

Solution Statement

This is an NLP sentiment analysis problem to understand user choices. This problem is a classification problem which can be solved using different classifiers (LGBM, Naïve Bayes, Random Forrest, Deep learning, etc.). The problem will be solved by using classification algorithms. It is multi class classification problem of human written text. The solution should be able to classify future incoming reviews into one of the five ratings with a certain level of accuracy.

Benchmark Model

The benchmark for this project has been taken from this Kaggle workbook. The benchmark is the accuracy scores for classifying the data using the following classifiers:

Classifier	Accuracy Score
LightGBM	59.48%
XGBM	58.06%
GaussianNB	50.74%
RandomForest	55.31%

The LightGBM classifier shows the best accuracy. The objective of this project would be to beat these accuracy scores by either optimizing one of the above-mentioned classifiers or using a different classifier.

Evaluation Metrics

The solution of this problem can be expressed in terms of accuracy of results predicted by the trained model. Accuracy is defined in as the ratio of correctly classified reviews and the total no. of reviews. Another interesting metric to track will be precision, since we would want to know how many of the reviews binned into one rating belong in that rating. For a two-bin classification problem, accuracy and precision are defined as follows,

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Project Design

The workflow which I will be following for this project is as follows:

- Data preprocessing – Understand and clean up the underlying data. This involves normalizing, tokenization, and splitting into test-train groups.
- Classifier testing – Testing different classifiers. As a starting step, I will limit myself to the classifiers considered in the benchmark.
- Classifier tuning – Hyperparameter optimization to achieve the highest possible accuracy score within reasonable model training times.
- Report preparation – Document the methods applied, and the results obtained during the project.