MLND Capstone Proposal

Domain Background:

Natural Language Processing has become the most talked domain in the recent years. As the applications like Text Classification. Topic Modelling, Machine Translation, Sentiment Analysis and Speech Recognition are emerging and using among all industries and become a vital field in the Information Technology. By saying this I am glad to be a part of NLP world and chosen as "Sentiment Analysis' as my MLND capstone project.

https://www.coursera.org/learn/python-text-mining

https://github.com/udacity/deep-learning/tree/master/sentiment-rnn

https://medium.com/udacity/natural-language-processing-and-sentiment-analysis-43111c33c27e

https://www.coursera.org/lecture/text-mining-analytics/5-4-how-to-do-sentiment-analysis-with-corenlp-jPNSG

Problem Statement:

The goal of the project is to classify the movie reviews based on the content. Here the content refers to description of the review. Here I have given with input text data and have the attached labels and this comes under supervised learning. Even though the use case seems solvable one but dealing with text is quite complicated compare to numerical data because of the ambiguity exits in here.

Datasets and Inputs:

Here the data represents the movie reviews with the relevant labels as positive and negative. The movie description considered as input and the labels as output for this use case. The total number of reviews in the dataset were 25K records which is quiet good enough to get a good model. The considered ratio of train, test and validation are 0.8, 0.1 and 0.1 respectively. Below is the link for the dataset. The target labels are balanced in the sense both the positive and negative labels has equal quantity with each of 12500 records.

https://github.com/udacity/deep-learning/tree/master/sentiment-network

Solution Statement:

Here the use case come under supervised learning. I would like to proceed by applying SVC/Gradient Boosting Classifier algorithm. By the way SVC and XGradient Boosting are simple/little complex and faster and works well for this kind of use case with the proper chosen of the feature engineering.

Here we try to assume that

Benchmark model:

As this is supervised learning approach trying to have a benchmark model will make sense to compare the accuracy. I would like to develop a benchmark model as SVC classifier with machine learning based feature engineering as TfIdf.

Evaluation metrics:

The evaluation metric considered here is the AUC score. Trying to maximise the precision/recall by minimizing the error term will make the model works better on the datasets. We get a better intuition when we plot an ROC curve. As part of training we will use grid-search techniques to tune the hyper parameters and chose the best parameters that give good AUC score.

Moreover by comparing the final model and benchmark model we can understand much better on both the models. We can compare both model performance metrics such as AUC score, training and prediction times.

Project design:

Let’s just mention the approach with the steps.

1. Data cleaning using NLP (Natural Language Processing) techniques. The steps under NLP are

    Text Cleaning (removing special chars, emoticons, symbols...),

    Tokenization, Parts of Speech Tagging, Lemmatization and stop words removal.

2. Feature Engineering using TfIdf/Embedding technique. Under Embedding will try applying Word2 Vec/Doc2Vec based methods.

3. Apply SVC/XGradient Boosting (XG Boost) algorithm.

4. Evaluate the each algorithm using Performance metrics such as AUC score.