DATS 6312 Project Proposal (Group 2)

Project Description:

The goal of this project is to estimate the Yelp star rating of a local business review based on the review text using NLP models with an emphasis on sentiment analysis. Yelp is a review-based website where people can discover and exchange information about local businesses. When consumers conduct a rapid search for businesses, they are more inclined to assess quality only on the star rating without reading the review language. As a result, our group is looking for probable connections between review wording and the corresponding star rating.

For this project, we will be utilizing the Yelp Review Sentiment Dataset. There are 560,000 training samples and 38,000 testing samples in all. The dataset we will be using is from the following source:
Dataset Link: https://www.kaggle.com/ilhamfp31/yelp-review-dataset

In this project we will be exploring NLP methods like pre-trained model, RNN model (LSTM). The baseline score will be determined using the conventional approach. Later, the models will be customized to increase the model's performance. Finally, the best-performing model will be displayed.

We will be using PyTorch packages such as torch, variable, get_tokenizer, nn, etc to implement the model. PyTorch is an open source framework and is capable of using GPUs effectively and efficiently. It is very similar to numpy hence it is easier to comprehend.

During the project, we will be working on tasks like cleaning the data, padding, packing, vectorising words, training a model and finally classifying the reviews accordingly. The performance of the model could be measured using a cross entropy metric.

Here is a rough schedule for completing the project-

Project Completion Steps	Duration
Data Preprocessing	11/20/2021
Modelling	11/27/2021
Evaluation	12/05/2021
Presentation Preparation	12/08/2021