Project 2

For due date see learn.bcit.ca

Yelp Rating Prediction using Sentiment Analysis

Deliverables

- The prepared report (in `.pdf` format) following the guidelines in `comp8085_report.tex`.
- The final source code implementation (preferably in `python`) which contains the implementation of your data preparation code as well as the training and inference code (which uses the data preparation procedures to transform only the input not the test set actual labels). Your submitted source code should provide the functionality to load the trained models and test them without going through training. Make sure somewhere in the report you mention how I can run your code in inference mode.
- Your held-out test set. I am not involved in data preparation step for this project, so make sure you submit your held-out test set on which you have reported your final results so that I can run your code in debug mode and try to replicate your results. **Do not submit the whole train set**, just the small portion you have held-out as your test set.
- A compressed file containing your pickled trained models so that they can be placed besides your code to run in inference mode.

- Make sure you submit all of these together at once since submissions in part will overwrite each other in the dropbox.
- Your submission must have 4 things **separately** in it. A `.pdf` file, a `.jsonl` file, a compressed file containing your source code and another compressed file containing your trained models. If your trained models are too big that cannot be uploaded to the dropbox, you may upload that single compressed file somewhere else and provide its download link in the textarea in the dropbox at the submission time.

Project Description

On Yelp.com, customers can rate a business in a range of 1 to 5 *stars*, and they can also provide a review (in text format) explaining their experience. Other optional rating fields that the customers can provide include: *useful*, *cool*, *funny*. These fields show the number of likes/dislikes that a post receives and for that reason are not bound to 5 (you may see values ranging from 1 to 1000+). The Yelp dataset includes a wide range of businesses such as restaurants, cafes, and bars. This project aims to develop a sentiment analysis model that can predict any of the mentioned values for a business on Yelp.com given the content of the review text.

Project 2

For due date see learn.bcit.ca

Acquiring the Review Data

As it is true for most of the datasets, the Yelp dataset is licensed by Yelp and you need to download this dataset from https://www.yelp.com/dataset. Once you click on **Download Dataset** button, you will be referred to a page that asks for *Your Name*, your *Email*, and your signature initials to allow you download it. Press **Download** button. Please read through the Dataset License **before using this dataset for any purpose** other than this course project, specially if the other purpose is non-academic. You will be redirected to the actual download page which allows you to download either the `JSON` dataset or the `Photos` dataset. You need to download the `JSON` dataset which will download one compressed file with the extension `.tgz` with the compressed size of 4.04GB which is uncompressed to 5 `.json` files[1] and 1 `.pdf` file (the *Dataset License* about which I talked earlier) totaling to the size of 8.65GB. Among the 5 compressed files, we are interested in the `yelp_academic_dataset_review.json` file. Here in one example line in the dataset:

```
{
    'review_id':"KU_O5udG6zpxOg-VcAEodg",
    'user_id':"mh_-eMZ6K5RLWhZyISBhwA",
    'business_id':"XQfwVwDrv0ZS3_CbbE5Xw",
    'stars':3.0,
    'useful':0,
    'funny':0,
    'cool':0,
    'text':"If you decide to eat here ...",
    'date':"2018-07-07 22:09:11"
}
```

Your task would be to receive the content of the `text` field and predict the values for `stars`, `useful`, `funny`, and `cool`. In **section one** of your report, provide information about the dataset you have acquired and your approach to extract evaluation data from it. You also need to settle on an evaluation metric to measure the performance of your models. Make sure you also talk about your evaluation procedure in this part of the report.

Part 1: Data Preparation

The dataset that you have acquired is quite raw. Design a procedure that can receive a dataset record (i.e. one json line like the one provided earlier) and gives back the processed record. Make sure this procedure is not dependant on the actual labels which means, if a dataset record is provided that does not contain the `stars`, `useful`, `funny`, or `cool` fields (which is true when the unseen data is fed to the procedure), the procedure must not crash.

[1] You can find more information in https://www.yelp.com/dataset/documentation/main about the content of each file

This procedure which can be implemented as a function or a class containing multiple functions will provide the input records to your sentiment analysis models. In the development procedure, to prepare the data, you may or may not choose to remove stop words, punctuations, and other irrelevant information, as well as, to convert the textual data to numerical vectors using techniques like *TF-IDF* (Term Frequency-Inverse Document Frequency) or *Count Vectorization*.

You may also choose to clean the training set by removing the records in which the important attributes `text`, `stars`, `useful`, `funny`, or `cool` are missing. **Please note that this cannot be done for the validation and test datasets!**

You may also use data visualization or any other techniques to get more familiar with the dataset records to do a better job in input preparation for your sentiment analysis models.

In the **Approach section** of your report, in one to two paragraphs, explain how you do data preprocessing and preparation on the training data json records and how does this differ to the case of input records in inference time.
Hint: I recommend reading the next section first and then getting started on this part since this part might need some specifications that would come from the model that is expected to use them.

Part 2: Model Selection and Experiment Design

Make sure your models are not using your validation set as training data. The validation and test sets **must remain the same** in all of your experiments.

Choose **different modelling techniques** (you need to choose one per group member and the techniques must be from different model classes) and in the **Approach section** of your report explain how each model works and how you will use each technique to perform sentiment analysis.
Your modelling technique selection should follow these rules:

- One technique must use a neural network design and you **cannot** use a pre-implemented model for this part. You may use `pytorch` library for designing the network. To get good results, I recommend doing an architecture that uses a pre-trained transformer model. Please note that this is a recommendation and you **don't** have to use transformer based models. However, they are very likely your best bet to get amazing results.

- One technique must use a probabilistic model and you **may** use different libraries to help implementing the model. You cannot take the whole model as a pre-implemented package, though.

- You are free to choose the other technique to your liking given that it is not a probabilistic model and not a neural network based model. You may use pre-implemented

`scikit-learn` package models for this one. I recommend a non-parametric model but you may choose other model classes.

- A model that reports an evaluation score below 0.5 does not count towards the modelling technique selection requirement.

- You are free to use the lab code and my released solutions to speed up this process. Just make sure wherever you use the lab code, properly cite and note it in your report.

- Since the models that you choose are different, your data preparation step might need to have different specialized data providers for each technique separately.

Train each model and using the held out validation set try to find the best settings for each. In your **Experiments and Analysis section**, explain the process of finding these best settings as well as the best evaluation scores on each of the four rating categories (`stars`, `useful`, `funny`, `cool`) for each model on the held-out test set. The results for all the techniques must be presented in a compact manner in a single table. Compare the performance of the models and try to figure out the strength and weaknesses of each model based on the category of inputs on which they normally make mistakes. This is considered the first experiment for each group member.

As suggested in the project guidelines (`comp8085_report.tex` file), design two other experiments per group member and conduct those experiments. Provide the results for each and analyze what you learned from their results in the report. For any two experiments in your submission, make sure they do not test the same idea.

Here are some experiment design ideas:

- Considering another model and comparing its performance with the ones we have already implemented.

- Preparing a separate hand annotated dataset and comparing the model results on our hand annotated real-world data.

- Performing ablation studies on the model performance, e.g. if we remove all the star ratings of 1 from everywhere, would our overall test score go up?

- If we remove the top k most frequent words from the text reviews (considering those as stop words), would our testing results change?

- ...

Be creative and curious with the experiments and come and talk to me if you need help.