Bridge-i2i Problem Statement

Group 14 -- H2_B21_14

Submission output files

The two output files, Output1 and Output2 are stored in the folder `Outputs` within the base directory.

Headline Generation Model

Of the four models we stated in our report for headline generation, we went with the ***Pegasus*** standalone model, by the virtue of better metrics across majority of the metrics.

Running Time

Task 1 takes 4.18 minutes (Theme Classification) Task 2 takes 4.25 minutes (Aspect based Sentiment Classification) Task 3 takes 33 minutes (Headline Generation) Preprocessing may take a variable amount of time, depending upon the number of requests to the Google Translation API by a given IP Address, and may be impacted by unavailibility of service due to high amount of API requests (in the scenario of running the code multiple times). {For stable runtimes of the translation process, the paid version of the API can be used. We, however, have used the open source version of the same, in accordance to the rules of the competition.}

Structure

The entire folder is arranged in the following manner:

1. Notebook Files:

There are five notebook scripts located in the base directory. A short description about them is as follows:

1. **Task1&2.ipynb**: Used to preprocess the dataset, and train distilBERT classifier and implement ABSA code for mobile brand and its corresponding sentiment analysis.
2. **Task3.ipynb**: Used to obtain metrics for different Headline Generation models, via Fine-tuned saved models.
3. **Pegasus_FineTune.ipynb**: Used to fine tune Pegasus for summarization task on the given dataset.
4. **T5_FineTune.ipynb**: Used to fine tune T5 model for summarization task on the given dataset.
5. **Evaluation.ipynb**: Used for evaluation of the testing data. Contains code for all three tasks.

5. Output Submission

Consists of two output files: Output1 and Output2. A description about them is as follows:

1. **Output1.csv**: Contains TextID, Predicted Labels, and generated headlines.

2. **Output2.csv**: Contains TextID, Predicted Labels and Extracted Brands and their sentiments.
3. **Preprocessed_Text.csv**: Consists of text data post preprocessing and translation. Located in the base directory

2. Saved Models

There are three folders containg saved models finetuned on the training dataset. They are:

1. **T5**: Contains T5 finetuned model for Headline Generation.
2. **DistilBERT**: Contains DistilBERT finetuned model for Theme Classification
3. **Pegasus**: Contains Pegasus finetuned model for Headline Generation.

3. Requirements

The requirements folder comprises of five `txt` files, containing required libraries version for each of the five notebooks mentioned above.

4. Development Data

Contains training data as provided under the problem statement.

5. Evaluation Datasets

Comprises of evaluation dataset as provided.