Qure.ai Assignment

1. Dataset

News from August 2009 are scraped from https://www.thehindu.com/archive/web. Rest of the news can be scraped by simply running `thehinducom_scraper.py` and wait longer. Scraper start from the beginning and scrapes news in chronological order.

There are 2370 news scraped within that period. After grouping similar categories into labels and removing news with vague categories, there are 2005 news left with 8 categories. Those categories are art, economy, education, industry, international, national, science & technology and sports.

2. Experimental Setup

Data split into training (70%), validation (10%) and test (20%) sets with stratified categories. Category distribution is preserved in those splits. Performance of the model is evaluated with accuracy, precision, recall and F1 score.

3. Model and Training

DistilBERT model is used as the classifier. It is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40%.

DistilBERT base uncased model is trained with cross-entropy loss and AdamW optimizer with 0.00001 initial learning rate. Learning rate is multiplied with 0.5 after 5 epochs without validation loss improvement. Batch size is set to 16. Total epochs are set to 100 and early stopping epochs are set to 15. Model converges around 70 epochs.

Training and validation losses during the training can be seen below.

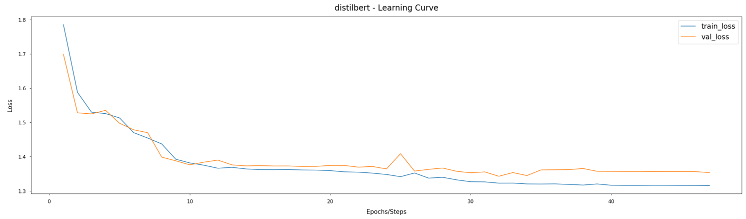

Figure 1: learning_curve

4. Results

Validation scores and test scores can be seen below.

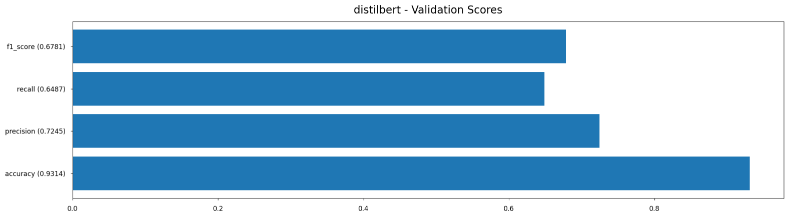

Figure 2: val_scores

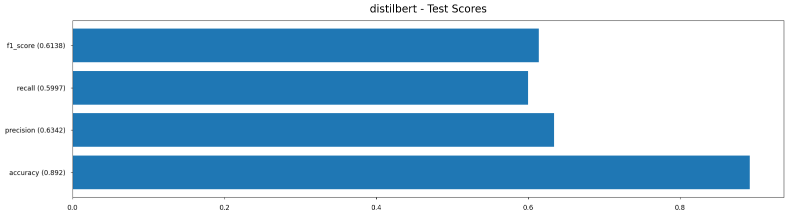

Figure 3: test_scores

5. Potential Improvements

- SotA text classification models (RoBERTa, DeBERTa and etc.) would probably score better than DistilBERT.
- Multilingual models might work as well since news content is related to India and there are reasonable amount of non-english words.
- Model head could be more sophisticated than a plain Linear -> Softmax layer.
- Validation should be done on specified iterations rather than after every epoch because transformer based models are very unstable.
- Models trained with larger max sequence length would probably score better.
- Tokenizer truncates news content after the sequence reaches specified max sequence length. This approach wastes lots of data because only the first 256 tokens are used for training. Every news content should be break into separate chunks of max sequence length tokens so all of the data can be utilized.

6. Web App

Web app can be launched by running `main.py` with `../config.yaml` and `deploy` arguments. This will load model with specified configurations and serve the web application on local host.