CS410 Final Project Documentation

squires4 | Samantha Squires

Contents

Project Summary

I trained a BERT text classification model to identify tweets as 'SARCASM' or 'NOT_SARCASM' using python/tensorflow/keras and Google Colab. I found this tutorial very helpful in the process of writing my code, and directly used some of the code provided in the tutorial–I've also noted this directly in my code in the relevant functions.

The code can be found in the Google Colab notebook in this repo (filename `TextClassificationCompetitionFinal.ipynb`). BERT was the first model I even attempted for this project, because I had heard of its success in text classification tasks. I didn't have much previous experience with tensorflow, so the main challenges that arose were understanding the formats and shapes required for model input and output.

To complete the project, I first utilized the Tensorflow BERT tutorial mentioned above to write a function for building a keras model using a pre-trained BERT model, and then a function to fit this model to the movie classification training set, then make predictions on the test data and return those predictions. Finally, I wrote a function to transform the model's predictions (log-odds values) to the format required by LiveDataLab, and save the result to a text file for downloading.

Finally, I wrote the main code pipeline, which loads the train and test data, preprocesses it by concatenating the 'response' and 'context' columns into a single feature and transforming the labels from 'NOT_SARCASM'/'SARCASM' into 0/1, and finally calls the functions that I wrote earlier to create, fit, and predict using the model. After the code is finished running, it outputs a file named `answer.txt`, which is in the proper format to be submitted via LiveDataLab.

How to Run the Code

All project code is located in `TextClassificationCompetitionFinal.ipynb`. Please run the code in **Google Colab** by following the steps below:

1. Download this repo.
2. Go to http://colab.research.google.com/ and click **Upload**, then select the downloaded file `TextClassificationCompetitionFinal.ipynb`. This should open the notebook in Colab.

The rest of these instructions are also included in the notebook itself, but for completeness:

3. In the Colab "Edit" menu, go to "Notebook Settings" and select "GPU" from the hardware accelerator dropdown.
4. Upload the train and test data files provided with the competition (make sure they're named `train.jsonl` and `test.jsonl`) by going to "Files" in the left-hand sidebar, clicking the upload icon,

and selecting `train.jsonl` and `test.jsonl`. These two data files are included in the repo, so you should already have them downloaded.

5. To run the code, select "Runtime" from the Colab menu and click "Run all".

The first few cells should run quite quickly: the final cell, which trains the model and predicts labels for the test data, takes much longer (in my experience, 10-12 minutes). After a minute or so, you should start to see output tracking the training progress of the model.

After the code finishes running, the output file, `answer.txt`, should be visible under "Files" in the left-hand sidebar. If you'd like to save this file, make sure to download it before the runtime disconnects.