Finetuning Language Models - Toxic Tweets

If you miss a milestone deadline you will be forfeited the corresponding points. Deadlines are quoted here as a week number - check your LMS about actual dates.

Introduction

Figure 1: toxic-comment

The internet has been converted from a tool offering unprecedented utility to mankind to a tool that can can destabilize societies and ven cause individuals to take their own lives cite.

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic

comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

Milestone 1: (Week 1, 20 points)

In this milestone you will learn the basics of docker and create a development environment. All AI and data science projects are developed in containers. Learn the basics of docker (the most conmmon container format) by watching the following video:

`{eval-rst} .. youtube:: pTFZFxd4hOI`

If you are on Windows you will need to follow these instructions and install Docker Desktop and WSL2.

Independent of your OS, you may want to use VS Code IDE if you have no IDE experience before. Ensure that you are able to debug code in your IDE. It must connect to the remote container.

Submit the github repository URL with a branch titled 'milestone-1' with the README.md file containing the installation instructions you followed and a screenshot of your docker container terminal prompt. Add as collaborator the TA.

Milestone 2: Sentiment Analysis App (Week 3, 20 points)

Merge the earlier branch into the main branch and create a new branch titled 'milestone-2. Do not delete the milestone-1 branch.

The purpose of this task is to take you through the process of creating a streamlit app. Streamlit is a python library that allows you to create web apps with minimal coding and deploy them in the cloud. This is a necessary step before you can develop your app for the more complex twitter use case.

After watching this video, you will be able to create a sentiment analysis app using Streamlit and various pretrained models. The pretrained models are available in the HuggingFace model hub.

`{eval-rst} .. youtube:: 8hOzsFETm4I`

Develop a streamlit application that allows the user to enter a text, select a pretrained model and get the sentiment analysis of the text. Use the HuggingFace transformers library just like in the video to do so.

Deploy the streamlite app in HuggingFace streamlit spaces after you create a free account.

Submit the github repository URL with a branch titled 'milestone-3' with the README.md file containing the link to the deployed HF space where the app must be clearly prepopulated with a sample text and the TA will only have to press the submit button for the app to display the sentiment result.

Milestone 3: Finetuning Language Models (Week 5, 40 points)

Merge the earlier branch into the main branch and create a new branch titled 'milestone-3. Do not delete the milestone-2 branch.

The following video will show you how to finetune a language model using the HuggingFace transformers library. Consult the HuggingFace documentation for more details.

```
{eval-rst} .. youtube:: GSt0O_-0ncQ
```

You will use this dataset to develop the required classifier. You're challenged to build a multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective's current models.

The classifier will be developed using a pretrained language model of your choice (e.g. bert, roberta, etc). You can use the GPU in Google Colab to do the training.

Submit the github repository URL with a branch titled 'milestone-4' with the README.md file containing the link to the deployed HF space where the app must be clearly prepopulated with a drop down menu to select the application filing number (or any unique identifier that will allow you to retreieve the patent sections you selected), the app must then show in minimally two text boxes the eg abstract and eg claims and the TA will only have to press the submit button for the app to display the patentability score.

Milestone 4: Documentation and Video Production (Week 6, 20 points)

Merge the earlier branch into the main branch and create a new branch titled 'milestone-4'. Do not delete the milestone-3 branch.

Document extensively both the code as well as the results (10 points).

Use google sites to create a landing page for your app (5 points)

Create a video that will demonstrate the app. The video should be no longer than 5 minutes and should be either uploaded to your youtube channel or inlcuded in the github repository as an mp4 (5 points).