ADVANCE Labs - Cyberbullying Detection Lab

The development of this document is partially funded by the National Science Foundation's Security and Trustworthy Cyberspace Education, (SaTC-EDU) program under Award No. 2114920.

1 Lab Overview

In this lab, you will learn about how AI/ML can be used to detect societal issues such as cyberbullying. Cyberbullying is bullying performed via electronic means such as mobile/cell phones or the Internet. The objective of this lab is for students to gain practical insights into online harassment such as cyberbullying, and to learn how to develop AI/ML solutions to defend against this problem.

In this lab, students will be given a starter-code. Their task is to follow the instructions provided in the Jupyter notebook, train an AI/ML model on the given dataset, evaluate their model, and deploy the model by testing it on their own samples. In addition to the attacks, students will also be guided to perform hyperparameter tuning to further improve the performance of their detection models. Students will be asked to evaluate whether their tuning effort improves their detection models or not. This lab covers the following topics:

- Detection of cyberbullying in tweets

Content Warning: This lab contains potentially triggering language and deals with difficult subject material. We minimized showing such language samples in this lab. They do not represent the views of the authors.

2 Lab Environment

This ADVANCE lab has been designed as a Jupyter notebook. ADVANCE labs have been tested on the Google Colab platform. We suggest you to use Google Colab, since it has nearly all software packages preinstalled, is free to use and provides free GPUs. You can also download the Jupyter notebook from the lab website, and run it on your own machine, in which case you will need to install the software packages yourself (you can find the list of packages on the ADVANCE website). However, most of the ADVANCE labs can be conducted on the cloud, and you can follow our instructions to create the lab environment on the cloud.

3 Lab Tasks

3.1 Getting Familiar with Jupyter Notebook

The main objective of this lab is to learn how AI/ML can be used to detect online harassment, such as cyberbullying. Before proceeding to that, let us get familiar with the Jupyter notebook environment.

Jupyter notebooks have a Text area and a Code area. The Text area is where you'll find instructions and notes about the lab tasks. The Code area is where you'll write and run code. Packages are installed using `pip`, and need to be preceded with a `!` symbol. Try accessing the lab environment for this task here.

The lab has three areas: one text area and two code areas. Follow the instructions for the three areas, fill the three areas with the instructed content and add a screenshot to your report.

3.2 Cyberbullying Detection

In this lab, you will develop AI to detect cyberbullying. You will use a dataset of tweets to train your AI model, evaluate the performance of your AI model, and then deploy it by running it against your own samples. You can access the lab by clicking here.

3.2.1 Datasets Selection

In this lab, we provide three datasets: formspring dataset, Davidson dataset and Founta dataset. You can edit the name of dataset in following code:

```
main_df = pd.read_csv('CyberbullyingLab1/formspring_dataset.csv', sep = '\t')
```

In this lab, you will be using the formspring dataset, a widely used dataset of cyberbullying texts. Run all the code until this part of the lab. Report the size of the training, testing and validation sets.

3.2.2 Preprocessing Data

Follow the instructions in the text areas and run the subsequent codes to preprocess your data, as follows. Here is a sample from the lab:

```
spacy_en = spacy.load('en_core_web_sm')
text = spacy_en("the cat sat on the mat.")
```

Add code to preprocess the following cyberbullying text, and include the generated tokens in your report: "`Harlem shake is just an excuse to go full retard for 30 seconds`". You can add a code block to run your code.

Proceed to further preprocess your dataset by using pretrained word embeddings.

3.2.3 Model Training

After you have preprocessed the dataset, the next task is to train the AI model. Follow the lab instruction to train the AI model. What is the final training accuracy that your model achieves?

```
print(Train Acc: {train_acc*100:.2f}%)
```

3.2.4 Model Evaluation

Now it is time to run your trained model on a test dataset. Recall that we have already partitioned the dataset into train, validation and test sets. Run your model on the test partition and report your results here. Use the evaluate function.

```
..., ... = evaluate(...) # complete this code
print(f'| Test Loss: {test_loss:.3f} | Test Acc: {test_acc*100:.2f}% |')
```

3.2.5 Deploy and Run Custom Samples

Download the samples file and use your model to detect the cyberbullying samples in this file. Report the samples that were detected as cyberbullying. Do you think your model is good enough? In this lab, you will later learn how to use hyperparameters to improve the performance of your model.

4 Submission Instructions

You need to submit a detailed lab report, with screenshots, to describe what you have done and what you have observed. You also need to provide explanation to the observations that are interesting or surprising. Please also list important code snippets followed by explanation. Simply attaching code without any explanation will not receive credits.