humane intelligence

Revontulet

Bias Bounty 2 – Counterterrorism
Humane Intelligence x Revontulet

Objective

Your task is to build an unsupervised model that can identify extremist content from the unlabelled sample image dataset provided. **This competition launches on Thursday September 26th at 9 AM ET, and closes on Thursday November 7th, 2024 at 11:59 PM ET.** We will be accepting submissions as of October 28th at 9 AM ET.

Note: to access the data for this challenge, **all participants must be 18+ and fill out this waiver form.**

Instructions

1. Model Development:
    a. Using the **training data** provided, develop an **unsupervised model to classify images** as hate or non-hate.
    b. Use the **test data** provided to validate your model
    c. Your model should output binary predictions:
        i. 1 for hate
        ii. 0 for non-hate
    d. Intermediate requirements
        i. Build an unsupervised machine learning model that groups unlabeled images into **2 clusters** to identify whether an image contains extremist content or not
    e. Advanced requirements
        i. **Building on top of the intermediate challenge**, create adversarial examples using the **test dataset** to test the robustness of your unsupervised model.
        ii. Explore different methods for generating adversarial examples with the provided image dataset that could potentially trick the model
        iii. Use your trained model to make predictions on the perturbed images

2. Starter Code to generate unique image identifiers:

```
import os

def generate_image_ids(image_folder):
    image_ids = []
    for image_file in os.listdir(image_folder):
        # Get the file name without the extension
        image_id = os.path.splitext(image_file)[0]
        image_ids.append(image_id)

    # Create a list of image IDs
    return image_ids


if __name__ == '__main__':
```

```
image_folder = 'path/to/image/folder'  # Update with your image folder path
image_ids = generate_image_ids(image_folder)
#print(image_ids)  # This will print the list of image IDs
```

3. Project Files:
    a. Predictions CSV:
        i. Your model should output a CSV file with predictions on the **test dataset**
        ii. The **first column in the CSV: "image_id"** is the name of the image file, and the **second column: "prediction_label"** is your model's classification of that image
            1. code for creating image_ids is below
    b. Model File:
        i. Submit the trained model file. This file should be in a "**.pkl" format *only,*** using the **Python Scikit-learn** library
    c. Inference Script in a **".py" format only**:
        i. Provide a script that can:
            1. Load the model file
            2. Load the sample dataset
            3. Generate predictions on the sample dataset
            4. Save predictions in the required CSV format
        ii. Ensure your script is **executable** and includes any **dependencies** and/or **instructions** needed to run it
    d. **Advanced-only**
        i. Submit folder with perturbed images in **".jpg" format only**
        ii. Submit an additional CSV file, with the predictions for the perturbed adversarial examples

4. Submission:
    a. Upload your project to GitHub as a private repo, **excluding the sample dataset**
        i. add @NicoleScientist as a collaborator to your private repo
        ii. Due to the sensitivity of the data, **participants are not permitted to upload the image data to GitHub**
            1. for advanced submissions only: upload your perturbed image dataset to your private repo
    b. We will begin accepting submissions on October 28th at 9 AM ET here.
        i. We are only accepting 1 submission per participant, please ensure you have all the **required files, and are happy with your solution**

Grading Outline

1. We will run your unsupervised model against our **holdout dataset (labelled), to compare** your model's predictions with the ground truth labels
2. Your submission **score will be based on accuracy** –calculated as the number of correct predictions divided by the total number of images

3. Since your unsupervised model may flip the binary labels, we will evaluate predictions assuming both possible label mappings (i.e., 1 = hate then 1 =non-hate).
    a. The higher score will be recorded