Project List: (100 points)
Each group of 5 students to choose *1* among the projects below to do:

IMPORTANT NOTE!! These data are copyrighted, and you should not distribute them or use them outside of this Fall2021 CS640 class.

(1) Gun Violence News **frame classification** using images. Use images *and* other information available in the image directory and data file to **predict the frame** of the news article (column Q3 Theme1 in the data file). The files are available in this directory.

Report the approach(es) you take and their 4-fold cross-validation accuracies for frame classification, and the per-frame precision and recall. You will be graded based on your (a) presentation and (b) write up *including* the link to your code (shared through GitHub/Google Colab) – particularly, how do you think through the problem and whether the approach you do make sense for the problem and the data, and the insights you obtain.

The current performances and approaches that have been tried can be found in this paper (EMNLP Findings, 2021) in Table 3. You don't have to match the performance, but if you outperform the current multi-class performance, we will award +20 bonus points for your project.

(2) Gun Violence News-Image **frame relevance**. Use images and other information available in the image directory and data file to **predict whether the image is relevant** to the frame of the news headline (column V3relevance in the data file). The files are available in this directory.

Report the approach(es) you take and their 4-fold cross-validation accuracies for relevance classification, and the per-class (relevant vs. not relevant) precision and recall. You will be graded based on your (a) presentation and (b) write up *including* the link to your code (shared through GitHub/Google Colab) – particularly, how do you think through the problem and whether the approach you do make sense for the problem and the data, and the insights you obtain.

The current performances and approaches that have been tried can be found in this paper (EMNLP Findings, 2021) in Table 4 and 5. You don't have to match the performance, but if you outperform the current performance, we will award +20 bonus points for your project.

(3) Demographic (**Age** (<21 and >=21) and **Race** (black, white, Hispanic/Latino, Asian) prediction of Twitter users. You are allowed to but are not required to predict the multiracial class. The data for this project is available in this directory. The training data is in the file *labeled_users.csv* in the directory. Description about this data can be found in the file *README* in the directory.

Your task is to use tweets of these Twitter users (available in the

Twitter_User_Handles_labeled_tweets.json file) to predict their **age** (<21 and >=21) *and* **race** (black, white, Hispanic/Latino, Asian) based on the training data available in the *labeled_users.csv* file. You can also use the profiles of these Twitter users (available in the *User demo profiles.json* file) to augment your training.

You can use the user's Twitter user_id in *labeled_users.csv* to obtain the user's last 100 tweets in the file *Twitter_User_Handles_labeled_tweets*.json

Similarly, you can use the user's Twitter user_id in *labeled_users.csv* to obtain their profiles (name, screen name, description, link to profile picture in the *profile pics.zip* file) in the file *User demo profiles.json.* You can use library such as ethnicolr to predict the race of the users from their names, and use these predictions to augment your training data for predicting race from tweets. You can also train a model for predicting race from the profile picture of the user or use dataset (and code) from VMER or UTKFace or others to train the model for predicting race from the profile picture. You can also use method described in this tutorial code to extract faces from the image and turn them into embeddings which you can use as features for your race prediction model.

You will build 2 models, one for predicting age and another for predicting race (you can use the same model for both tasks).

Report the approach(es) you take for each of this task and their 5-fold cross-validation accuracies for age classification and for race classification (on the labeled data where age is not null and race is not null, respectively), and the per-class precision and recall (<21 or >=21 for age, and black/white/Hispanic-Latino/Asian for race).

You will be graded based on your (a) presentation and (b) write up *including* the link to your code (shared through GitHub/Google Colab) – particularly, how do you think through the problem and whether the approach you do make sense for the problem and the data, and the insights you obtain.

The current performances and approaches that have been tried can be found the file *stats_race_&_age.pdf*. The 5-fold cross validation accuracy for the 4 classes of race is 0.67. You don't have to match the performance, but if you outperform the current multi-class performance for *either* age or race, we will award +15 bonus points each (so +30 bonus points if you improve both) for your project.

(4) COVID-19 Instagram posts **emotion detection** (**anger**, **fear, joy, and sadness**) in relation to images of **East Asian people**. In this project, you will first predict the emotion in the Instagram posts that are in the file (*Labeled_instagram_posts_related_to_covid*.xlsx) in this directory. The posts are not annotated for emotion, but they are annotated for whether the image accompanying the post contains image of East Asian person. The description of the annotation can be found in the *README_instagram_label.docx* file in the directory (particularly, you only need to use Q5a that annotates whether the image contains an East Asian person—i.e., label 1). The images can be found by linking *imagename* column in this file to the picture in the

images.zip file in the directory.

You need to build a model for detecting emotion in the Instagram post. You will use emotion Twitter dataset here (without intensity label) to train the model for predicting emotion.

Report the approach(es) you take for each of this task and your multi-class accuracy and per-class precision and recall for each emotion class (fear, anger, joy, and sadness) **on the development set**. You also need to apply your emotion detection model on the Instagram posts in *Labeled_instagram_posts_related_to_covid*.xlsx file and submit the file containing the predictions with your final project report. **Note**: For this project, the write-up has to include your analysis if there is a correlation between the presence of emotion (anger/fear) with the presence of East Asian person in the image (you can use the emotion predictions you have and the labels: label 1 in Q5a in *Labeled_instagram_posts_related_to_covid*.xlsx to measure this correlation).

You will be graded based on your (a) presentation and (b) write up *including* the link to your code (shared through GitHub/Google Colab) and (c) your emotion predictions. Particularly, you will be graded on how do you think through the problem and whether the approach you do make sense for the problem and the data, and the insights you obtain. Bonus +10 points for the team that has the highest multi-class performance on the development set (provided there are more than 1 team choosing this project).

Bonus: another +20 points if you also build a model for detecting whether there's an Asian person in the image. You can use method described in this tutorial code to extract faces from the image and turn them into embeddings. You can then use the embedding of the extracted face to predict whether the face is that of an East Asian person or not (i.e., binary prediction). You can use labels from the training data (i.e., Q5a in *Labeled_instagram_posts_related_to_covid*.xlsx) to train this model and report 5-fold cross validation accuracy and precision and recall of your model for predicting that there is an East Asian person on the data. ***Or***, you can train a model for predicting if a face is that of an East Asian person using labels from other dataset (and code) from VMER or UTKFace *and then* use the trained model to make prediction and report the accuracy and precision and recall of your model for predicting that there is an East Asian person on the labeled data in *Labeled_instagram_posts_related_to_covid*.xlsx.