Final Project Deliverables due Tuesday, April 20th, 2021

Vaed Khurjekar, Paul Wei Zou, Paula Scanlan, Anna Callahan, Sophie Perez, Joao Leite

Raw data input

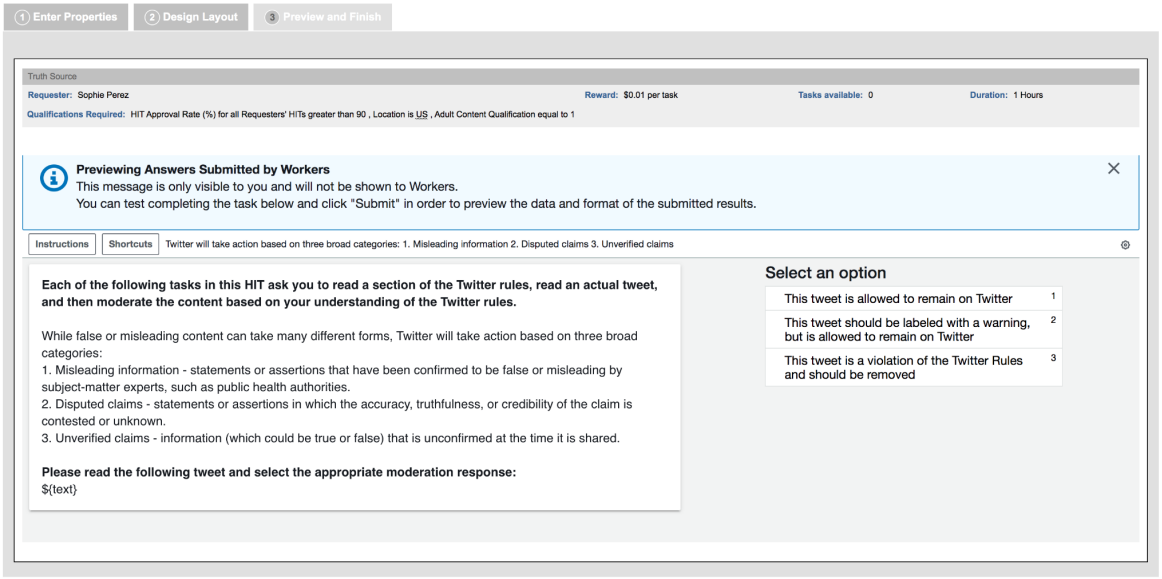
The goal of this project is to leverage the wisdom of the crowd to moderate content on Twitter. The input data will consist of tweets pertaining to COVID-19 and U.S. elections. We need to collect a sample of 100 tweets with roughly 50 pertaining to each topic. Within these subcategories, we have collected tweets in the following categories:

- **Control** (at least 12 tweets per topic, 24+ tweets total)
  - Permissible content according to the Twitter Rules
  - Go onto twitter and select tweets related to either COVID-19 or U.S. elections that adhere to the rules and don't fall into any of the below categories.
- **Unverified Claims** (at least 12 tweets per topic, 24+ tweets total)
  - Information (which could be true or false) that is unconfirmed at the time it is shared and doesn't carry a direct or significant propensity for harm
- **Disputed/Misleading Claims** (at least 12 tweets per topic, 24+ tweets total)
  - Disputed content or misleading content without severe consequences
- **Prohibited Content** (at least 12 tweets per topic, 24+ tweets total)
  - Misinformation or disinformation with direct and severe consequences

All tweets are contained in this spreadsheet.

Sample input/output from your QC module

Workers need to have HIT an approval rating of greater than 90%. Workers must be only in the United States. Each HIT will consist of one Gold Standard Question as well as one tweet each from the control, unverified, disputed/misleading, and prohibited categories. If a user fails the Gold Standard then all of their data will be rejected. The following is a screenshot of an example task:



Sample input/output from your aggregation module

Each tweet will be checked by 4 individual workers. Workers will be asked for their political affiliation as the first question of the HIT, and we hope to gain an even amount of workers from both major parties (Republicans and Democrats). We will use the simple majority aggregation technique as one component of this analysis. We will create a collab notebook containing a function similar to the simple majority function used in the previous homework to accomplish this. We will also write a few functions capable of extracting the following insights from our data:

- Was there a correlation between political parties and responses?
- Were users capable of making moderation decisions that aligned with Twitter's decisions?
- Were some categories easier or more difficult for workers to moderate?

1. Write the code to design the HITs according to what we decided for #2. - Anna, Sophie, and Paula

   Example HIT
   Task 1: Read the following excerpt of the Twitter Rules and select the appropriate moderation response. (this is the GSQ, so it needs to be a straightforward answer)
   Task 2: Control
   Task 3: Unverified
   Task 4: Disputed/misleading tweet
   Task 5: Prohibited

2. Write code that will address quality and analyze data (how do we balance out the multiple people that look at each tweet? Are we looking for a specific answer? Do we reject people, or take specific answers out of our data for inaccuracies?

   https://colab.research.google.com/drive/1TblonVwx0-uY9O8N5BvwbCGaoEJIMXOD?usp=sharing

   Our first function *preselect_worker* conducts two main quality control checks and serves as the first boundary for our HIT output data. First we will check to ensure that we only include those who answered. We make sure that any null or missing answers will be accounted for. Next we want to make sure that we only continue to use workers that show a basic competency of the instructions on a very black ad white moderation example. We will denote a "pass" score of 1 if the corresponding worker id has answered correctly and a 0 otherwise. Moving forward, we will build on this by only aggregating and analyzing the results of those that have passed this checkpoint.

$15 per student * 6 group members = $90 budgeted for this project
We may have to run some of this again, so we're going to say we're budgeting $60 for the first HIT