NLP PROJECT DATA SHARING AGREEMENT

Purpose of Agreement

The purpose of this agreement is to ensure provided Tweets are not distributed and remain within the educational bounds of this NLP project. These Tweets were downloaded using the Twitter API for educational and non-commercial research purposes. By signing this agreement, the student agrees to utilize provided Tweets solely for this NLP project, and will delete these Tweets upon the completion of the semester.

Description of Data

Tweets and Candidate misinformation targets will be provided in two separate files. The first file will be in jsonl format containing Tweet IDs and the textual content of the tweet. An example of this file follows:

{"id": "1340150211423952896", "text": "Covid Vaccine is the new version of populationg control to combat climate change."}

{"id": "1340070783536812034", "text": "Do not take the COVID-19 vaccine. You will never be the same."}

...

The candidate misinformation targets file will also be in jsonl format and will contain Tweet ID / misinformation target ID pairs which will be annotated by students. An example of this file follows:

{"tweet_id": "1340015106206392321", "m_id": "9"}

{"tweet_id": "1340570906662322176", "m_id": "2"}

...

Students will be expected to submit annotations for each of these Tweet ID / misinformation target ID pairs in a jsonl file with the following format:

{"tweet_id": "1340015106206392321", "m_id": "9", "m_label": "agree"}

{"tweet_id": "1340570906662322176", "m_id": "2", "m_label": "agree"}

...

Where "m_label" will take one of the following values:

["agree", "disagree", "no_stance", "not_relevant"]

Each m_id represents one of many misinformation targets, which will be provided in a json file with the following format:


```
{
  "1": {
    "title": "mRNA alters DNA",
    "text": "RNA alters a person's DNA when taking the COVID-19 vaccine.",
    "alternate_text": "RNA alters a person's DNA when taking the coronavirus vaccine.",
    "source": "https://en.wikipedia.org/wiki/Misinformation_related_to_the_COVID-19_pandemic#Vaccine_misinformation"
  },
  "2": {
    "title": "Infertility",
    "text": "The COVID-19 vaccine causes infertility or miscarriages in women.",
    "alternate_text": "The coronavirus vaccine causes infertility or miscarriages in women.",
    "source": "https://en.wikipedia.org/wiki/Misinformation_related_to_the_COVID-19_pandemic#Vaccine_misinformation"
  },
...
}
```


Data Access

All training Tweets will be provided to all students on eLearning, but each student will only be responsible for annotating their 200 Tweet / Misinformation target candidates. These 200 candidate pairs per student will be provided in five files on eLearning called X1_train_candidates.jsonl, X2_ train_candidates.jsonl, …, X5_train_candidates.jsonl, where each file contains 200 candidate pairs. The student with the responsibilities of X1 from the

project proposal will annotate the X1_ train_candidates.jsonl, X2 will annotate X2_train_candidates.jsonl, etc. These annotated pairs will be submitted on eLearning. After the annotation submission deadline we will release additional annotated training examples, to raise the training data size from 1,000 total pairs (200 per student in each group) to 2,000 total pairs. Each group is responsible for merging their team's annotated pairs with the additional 1,000 pairs for training and evaluating their systems.

A final release on eLearning of test Tweets along with test candidate Tweet / misinformation target pairs will be provided to students at the end of the project on which they will be expected to perform stance detection. Their predicted stances will be submitted on eLearning in the same format as annotations (jsonl, with tweet_id, m_id, and m_label) and will be evaluated on the test labels. Performance of each team will be released on a leaderboard on eLearning. Performance will be evaluated with Precision (P), Recall (R), and Macro-F1.

Under NO CIRCUMSTANCES should Tweet IDs or text be shared from the data provided within this project. Tweets and annotations should be deleted upon completion of the project. Any questions regarding the file format, annotation process, or evaluation protocol should be emailed to the TA.

Signature

Chaoran Li

Signature

Chaoran Li

Printed Name

cxl190012

NetID

04/03/2021

Date