ECE 364 Project Option #2: Identification of Informative COVID-19 English Tweets

Spring 2025

1 Project Background

This project is based on the WNUT-2020 Competition[1]. The objective of the task is to identify whether or not a given tweet related to COVID-19 is informative (Fig. 1). Such informative Tweets provide information about recovered, suspected, confirmed, and death cases as well as location or travel history of the cases. The dataset can be downloaded from the competition' s GitHub repository[2]. The dataset contains three sets – training (`train.tsv`), validation (`valid.tsv`), and test (`test.tsv`). You should only use the training set to train the model. The validation set can be used to tune the hyperparameters or select the best checkpoint in training. The final scores must be reported on the test set.

The dataset comes in the form of `.tsv` files. Each file contains three fields:

1. `Id`: ID assigned to a tweet.
2. `Text`: The content of the tweet.
3. `Label`: Either `INFORMATIVE` or `UNINFORMATIVE`.

This is a binary classification problem. We want to classify each test example in one of two classes. For evaluation, please generate **prediction.csv** file in the following format. Accuracy score will be calculated by kaggle as the evaluation metric.
Note: To ensure effective evaluation, head and content of prediction.csv file must exactly match the format, case of text matters.
You can use the " submit prediction" button at the upper right corner in kaggle page to submit, remember to write names of all group members in the description. After submitting, you need to go to the " Submissions" tag and press the " select" box after your best submission, so your submission will be shown on the leaderboard.
Note: use the original id provided by the test.tsv, the ids in the table are only for reference

Id	Label
1	INFORMATIVE
2	INFORMATIVE
3	UNINFORMATIVE
...	...

Table 1. Format of **prediction.csv**

[1] `https://competitions.codalab.org/competitions/25845`
[2] `https://github.com/VinAIResearch/COVID19Tweet`

2 Deliverables

The following are the deliverables of this project:

- A binary classifier to classify each tweet in the test set as `INFORMATIVE` or `UNINFORMATIVE`.
- The number of parameters in the model must not exceed 15 million.
- You are free to use any publicly available model (pre-trained or otherwise) with or without augmentation, but it is not a requirement. You can also augment the data as you see fit.

As a starting point, check HuggingFace[3] for readily available pre-trained language models. Try models like Tiny-BERT. The idea is to get some hands-on experience with training models, so don't worry too much about getting a very high accuracy score. Any reasonable model is fine.

3 Submission

1. Submit all your code, including training and evaluation, as a `.zip` file.
2. Submit **prediction.csv** file to Kaggle for scoring and evaluation
3. Submit a 2-page report (1-inch margin, 12-point font) and include
   - Your approach, model, and any other design choice.
   - Hyperparameters that you used for training.
   - Training and test results.
   - Any other interesting details about the approach or model.

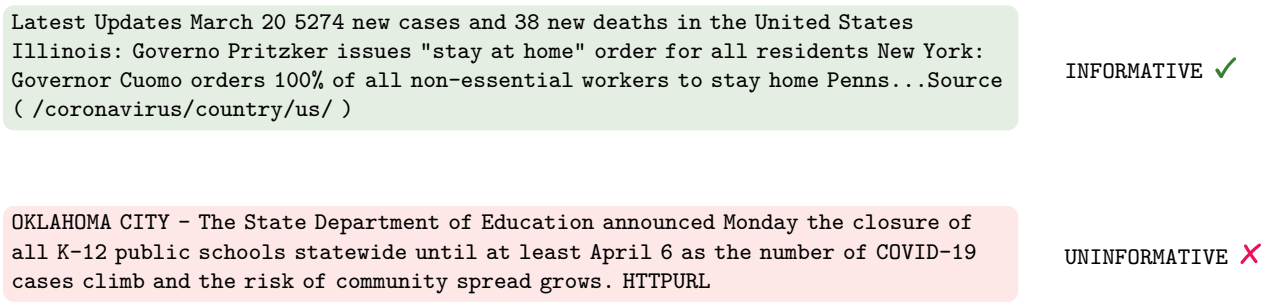

Fig. 1. An informative and uninformative example from the training set.

[3] `https://huggingface.co`