Disaster Tweets Classification by BERT and Its Variants

Group 3: Xuefang Hu, Anupam Kumar, Changchang Liu

1. Description

Many agencies, for example, news agencies and disaster relief organizations are now interested in monitoring Twitter as it becomes more and more important in people's communications everyday, and even in emergency times. However, the words from smartphones can be very ambiguous. Things such as metaphor, equivocation and implicitness exist in human languages, and they are much unclear to machines compared to humans. We are going to build a model that can predict which Tweets are about real disasters and which ones are not.

2. Dataset

We will use the dataset https://www.kaggle.com/c/nlp-getting-started/data from Kaggle, where each row contains the text of a tweet, a keyword from the tweet and the location that the tweet is sent from. However, the keyword and the location might be null in some rows. And we want to predict whether a given tweet is about a real disaster or not.

3. Methodology and Expected Results

The major outside tools that we plan to use are Scikit-learn and Tensorflow, as well as matplotlib for some visualization. We plan to use the BERT model as well as one of its variants (e.g. ALBERT/RoBERTa) to do the prediction and compare their training time as well as performance.

The main results we want to have are the labels showing whether a tweet is a real disaster (label = 1) or not (label = 0). We also hope to do some visualization, for example, show the top locations related to disaster tweets, plot the training loss and accuracy, and plot the pooled outputs and sequence outputs of the classification.

4. Timeline

Week 9: Clean and preprocess the data.

Week 10 & 11: Implement the two models and train them.

Week 12: Test and do final minor adjustments. Visualize the results.

Week 13: Write the report and record the final presentation.

5. Responsibilities

There are three members in our team: Xuefang Hu, Anupam Kumar and Changchang Liu. Each member is going to contribute nearly evenly in this project. Xuefang will be mainly responsible for cleaning the data and evaluating the models on the test set, Anupam will mainly implement and train the BERT variant model and Changchang will mainly do the BERT model and the visualization of the results for both models.