Tool: Google Collab (easy to collaborate, free and get use GPU)

1. Reading in data:
   1. Load “train.json” file
2. Data processing
   1. Data Cleaning
      1. Lowercasing
      2. Remove punctuation
      3. Remove special characters and numbers
      4. Remove stopwords
      5. Check missing data and remove duplicated rows
      6. Tokenization
   2. Convert raw data into specific data formatting
      1. NLTK
3. Model Selection
   1. RNN (TBC) - Stacy
   2. CNN - Jiahui
   3. BERT - Lily
      1. [bert-base-multilingual-uncased-sentiment](https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment) or
      2. bert-base-uncased
4. Training
   1. Try different model parameters
5. Prediction
   1. Report results in “submission.csv”

|  |  |  |
| --- | --- | --- |
| **Deadline** | **Person-in-Charge** | **Task** |
| 18 Oct | Lily | Reading in data |
| 20 Oct | Stacy | Data Cleaning:   * lowercasing * remove punctuation |
| 22 Oct | Lily | Data Cleaning:   * Remove special characters and numbers * Remove stopwords |
| 24 Oct | Jiahui | Data Cleaning:   * Check missing data and remove duplicated rows * Tokenization |
| 30 Oct | All | Convert raw data into data formatting using NLTK |
| 10 Nov | All | Model Selection, Training and Prediction |
| 16 Nov | All | Report |