README

Group 13 - Toxic Spans Detection

December 14, 2020

1 Model Link

Link : semi_sup-6.pt

- ’semi_sup-6.pt’ is our best performing trained model which has been used to predict the offsets provided in the next section.

2 Result Link

Link : predicted.csv

- ’predicted.csv’ file contains 2 columns, ’spans’ and ’text’.
- ’spans’ column contain the predicted offsets of the toxic characters corresponding to the ’text’ entry.

3 Colab Links

3.1 Training File

Link : Final_Train_Code_BERT

- This is the script we have used to train our best performing model
- To run the script, you need to provide

  1. train_path : path to .pkl file of preprocessed training data (train.pkl - section 4)
  2. val_path : path to .pkl file of preprocessed validation data (val.pkl - section 4)
  3. aug_path_1 : path to .pkl file of the non-annotated data whose labels we have predicted using our model for Semi-Supervised Learning. (aug.pkl - section 4)

- While training the model, we evaluate our performance using token level F1 score. To check the performance of trained model on validation dataset in terms of competition metric, pass link to the trained model and 'dev-final.csv' to the prediction code below and evaluate using the evaluation script.

3.2 Prediction File

Link: Finalprediction.ipynb

- This script takes the input file 'test.csv' provided by TAs and is used to make predictions by our model(section 1).
- To run the script,you need to provide
  1. 'test_path' : the path where 'test.csv' is located.
  2. 'model_path' : the path to our model provided in section 1.
  3. 'save_path' : the path where you want to save the final predictions in csv format(predicted.csv).

3.3 Evaluation File

Link: Evaluation.ipynb

- This script is the metric calculator(f1 score) given the true labels and predicted labels.
- To run the script, you need to provide
  1. 'test_path' : the path to test split(which contain true offsets in column 'spans'). We have assumed this file is similar to train.csv and dev.csv split provided to us.
  2. 'pred_path' : the path to the predicted.csv provided in section 2.

4 Preprocessing and other files

4.1 Preprocessing

Link : Final_Preprocess.ipynb

- This is the script we have used to preprocess our training and validation data splits(provided by the TA's) that are used while training.
- To use this script you will need to provide
  1. file_path : path to .csv file containing the original data splits
  2. save_path : path to .pkl file in which the preprocessed data will be saved

4.2 Semi-Supervised Learning Prediction

Link : Final_SSL_Data.ipynb

- This script is used to tokenize the unannotated data taken from Civil Comments Dataset and make predictions on it using our model. The predicted labels and tokenized data are saved to train further iterations of Semi-supervised learning models.
- To run this you need to provide
  1. path : path to the civil.csv file provided below
  2. model_path : path to the base model used make predictions on the non-annotated data
- If you want to use this notebook to recreate the aug.pkl file, use the semi_sup-5.pt model linked in section-4.

4.3 Data files

1. trainfinal.csv : This was the train split provided to us by the TAs.
2. devfinal.csv : This was the dev split provided to us by the TAs.
3. test.csv : test split (without labels) provided by the TAs.
4. civil.csv : File which contains the civil comment dataset used in SSL prediction notebook.
5. train.pkl : It is obtained by preprocessing the trainfinal.csv using Final_Preprocess.ipynb. Our data is trained using this file.
6. val.pkl : It is obtained by preprocessing the devfinal.csv using Final_Preprocess.ipynb.
7. aug.pkl : File obtained after predicting on the non-annotated data from Civil Comments Dataset. Obtained using Final_SSL_Data.ipynb script.
8. semi_sup-5.pt : Link to model from previous iteration of Semi-supervised learning. To be used in Final_SSL_data.ipynb to recreate the aug.pkl data file.