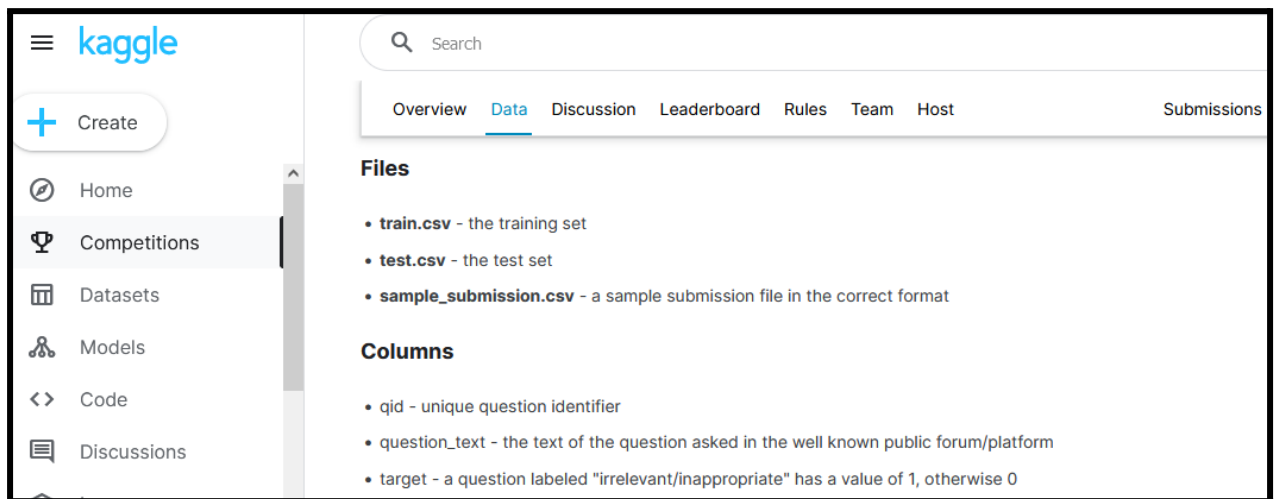




3. You can see the train and test set from the Data section. We are providing a Colab Notebook ([https://drive.google.com/file/d/1URA0r2Ku82IFtgprmKzD8DD4jeo\\_Jdyj/view?usp=sharing](https://drive.google.com/file/d/1URA0r2Ku82IFtgprmKzD8DD4jeo_Jdyj/view?usp=sharing)) which contains the instructions for downloading the train and test data. However, the train, test data and submission files can be downloaded from Kaggle also, but we recommend downloading train data, test data and sample submission file by executing cells in the Colab Notebook.



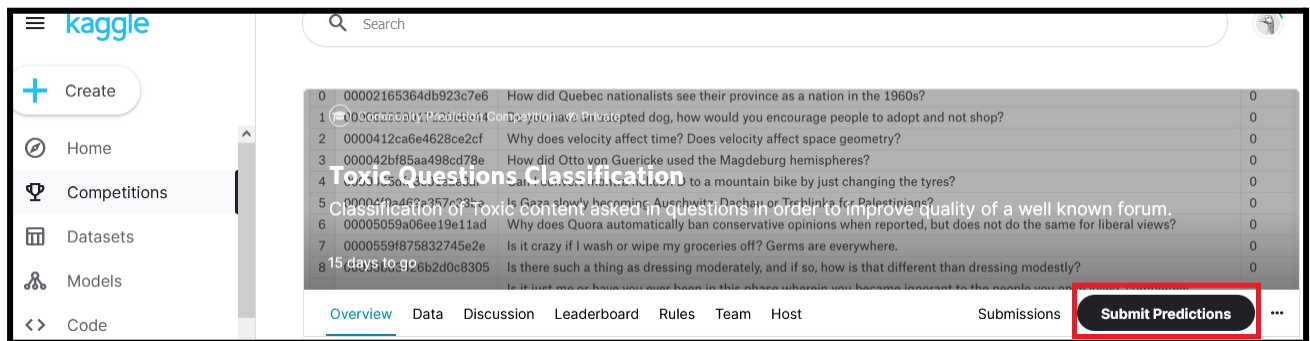
You will obtain

- A. Train directory that contains approximately 1044897 questions selected from each of 2 classes toxic/inappropriate (also known as irrelevant) and relevant/appropriate.
- B. Test directory that contains 261221 questions selected from each of 2 classes toxic/inappropriate (also known as irrelevant) and relevant/appropriate.
- C. sample\_submission.csv, which is a template for submission or predictions.

**Note:** Do follow the exact order of question denoted by qid as given in the test data while making predictions in the sample\_submission.csv file).

4. Build an RNN (Long Short Term Memory/Gated Recurrent Unit) or a Convolutional Neural Network (CNN) or BERT (with/without 'attention') using either Keras or PyTorch deep learning libraries on train and test dataset samples and perform classification of approximately 261221 questions asked on a well known public form as 'toxic/inappropriate' questions or 'relevant/appropriate' questions. Please predict the results to sync with sample\_submissions.csv. You are also welcome to use pre-trained models such as BERT (with 'attention' or without 'attention') to train and evaluate the test data.

5. sample\_submission.csv contains the format and header File Name and Target). Prepare your predictions in sample\_submissions.csv format. Submit it by clicking on the Submit predictions tab as shown below.



6. After successful submission, you will see the F1 score generated and an updated leaderboard position. The test set is split into a public set (80%) and a private set (20%). All your submissions are evaluated on both parts of the test set. By default we take your best submission on the public set for rankings. The public leaderboard is the ranking of all participants' submissions on the public set. It is actualized after every submission. The private

leaderboard is the rankings of participant's submission on the private set, it is hidden from you and used for evaluation.

- The public leaderboard position will be updated as you submit the best predictions. The scores of the different teams will be displayed in real-time on the public leaderboard.
- The final score is calculated on the private leaderboard.
- Maximum 20 submissions are allowed per day (IST time) per team -
- The public leaderboard will be active till 11:59 PM, 28th October, IST -
- The private leaderboard will be visible after 11:59 PM, 29th October, IST

## 7. Embeddings

Datasets other than the ones provided on Kaggle are not allowed for this competition. However, you can make use of the following word embeddings along with the train and test datasets that can be used in the models. These are as follows:

- **GoogleNews-vectors-negative300** - <https://code.google.com/archive/p/word2vec/>
- **glove.840B.300d** - <https://nlp.stanford.edu/projects/glove/>
- **wiki-news-300d-1M** - <https://fasttext.cc/docs/en/english-vectors.html>