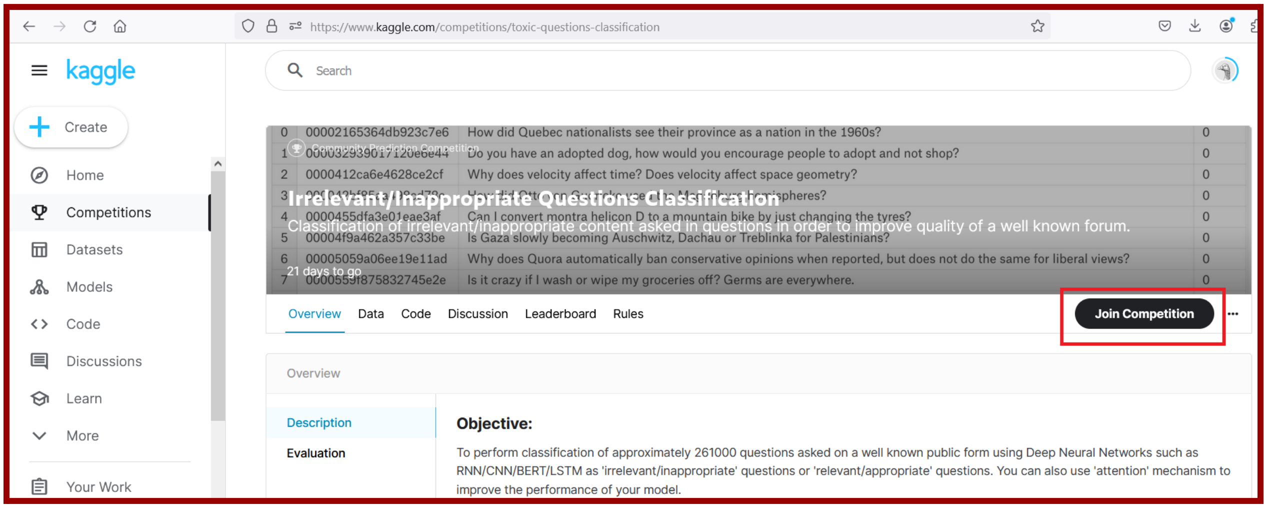
ReadMe

DLFA Module 03 – Irrelevant/Inappropriate Questions Classification

Team Creation and Prediction Submission in Kaggle

1. If you do not have an account with Kaggle, you can use your google account to **sign in** to Kaggle or create one for yourself or one on behalf of your team comprising four to five members.

2. Go to this in-class Kaggle Competition (https://www.kaggle.com/t/bde6f23028154933a99e4b4ca8a3dff2) and **accept the rules** under the "Rules" tab



0	00002165364db923c7e6	How did Quebec nationalists see their province as a nation in the 1960s?	0
1	000032939017120e6e44	Do you have an adopted dog, how would you encourage people to adopt and not shop?	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity affect space geometry?	0
3	000042bf85a[illegible]	How did Otto von Guericke use the Magdeburg hemispheres?	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain bike by just changing the tyres?	0
5	00004f9a462a357c33be	Is Gaza slowly becoming Auschwitz, Dachau or Treblinka for Palestinians?	0
6	00005059a06ee19e11ad	Why does Quora automatically ban conservative opinions when reported, but does not do the same for liberal views?	0
7	0000559f875832745e2e	Is it crazy if I wash or wipe my groceries off? Germs are everywhere.	0

Irrelevant/Inappropriate Questions Classification

Classification of irrelevant/inappropriate content asked in questions in order to improve quality of a well known forum.

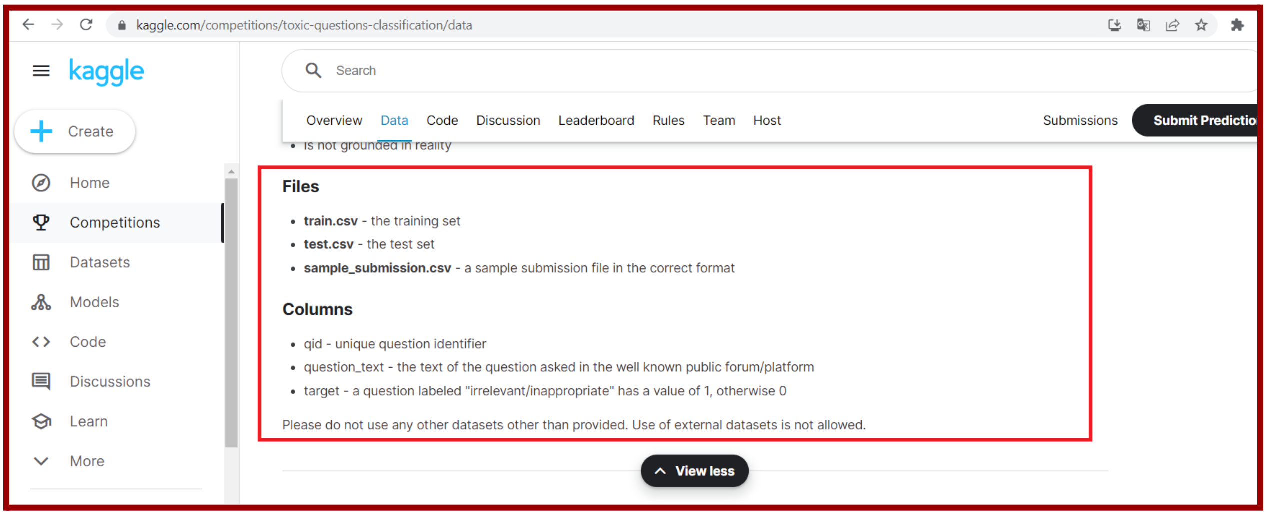
21 days to go

Objective:

To perform classification of approximately 261000 questions asked on a well known public form using Deep Neural Networks such as RNN/CNN/BERT/LSTM as 'irrelevant/inappropriate' questions or 'relevant/appropriate' questions. You can also use 'attention' mechanism to improve the performance of your model.

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/1gTxZR96YQoejhK_uli2-OsI06h__fLPn/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.



- is not grounded in reality

Files

- **train.csv** - the training set
- **test.csv** - the test set
- **sample_submission.csv** - a sample submission file in the correct format

Columns

- qid - unique question identifier
- question_text - the text of the question asked in the well known public forum/platform
- target - a question labeled "irrelevant/inappropriate" has a value of 1, otherwise 0

Please do not use any other datasets other than provided. Use of external datasets is not allowed.

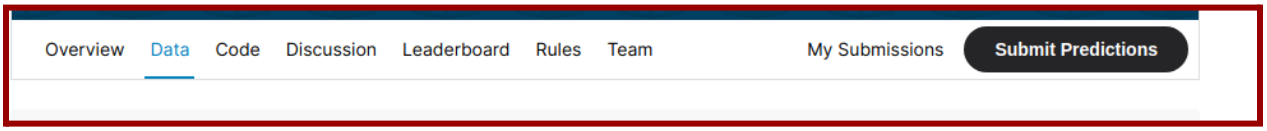
You will obtain

A. Train directory that contains approximately 1044897 questions selected from each of 2 classes irrelevant/inappropriate (also known as toxic) and relevant/appropriate.
B. Test directory that contains 261221 questions selected from each of 2 classes irrelevant/inappropriate (also known as toxic) 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 'irrelevant/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, 17th June, IST
- The private leaderboard will be visible after 11:59 PM, 17th June, 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