Multilingual Toxic Comment Classification

Problem statement

The goal of this project is to identify toxicity in online conversations, where toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion. It only takes one toxic comment to sour an online discussion. If these toxic contributions can be identified, we could have a safer, more collaborative internet. The challenge here is to build multilingual models with English-only training data.

Background

With the rise of social media platforms, online communication and discussion has become an essential part of people's internet experience. Unfortunately, communications online can often be quite rude and vulgar. The threat of abuse, bullying and harassment online means that people stop expressing themselves and give up on seeking different opinions.

This issue has triggered both the industrial and research community in the last few years. Whilst there are several attempts being made to identify an efficient model for online toxic comment prediction, the main area of focus is definitely on using machine learning/deep learning models that can identify toxicity in these conversations. If toxicity online could be identified, we could have a safer, more collaborative community online.

The challenge in the recently held Kaggle competition 'MultiLingual Toxic Comment Classification' was to create a model that predicts the toxicity of various comments that are multilingual in nature. They provided a training set that had various files from their previous competition, a validation set and a test set.

Methodology

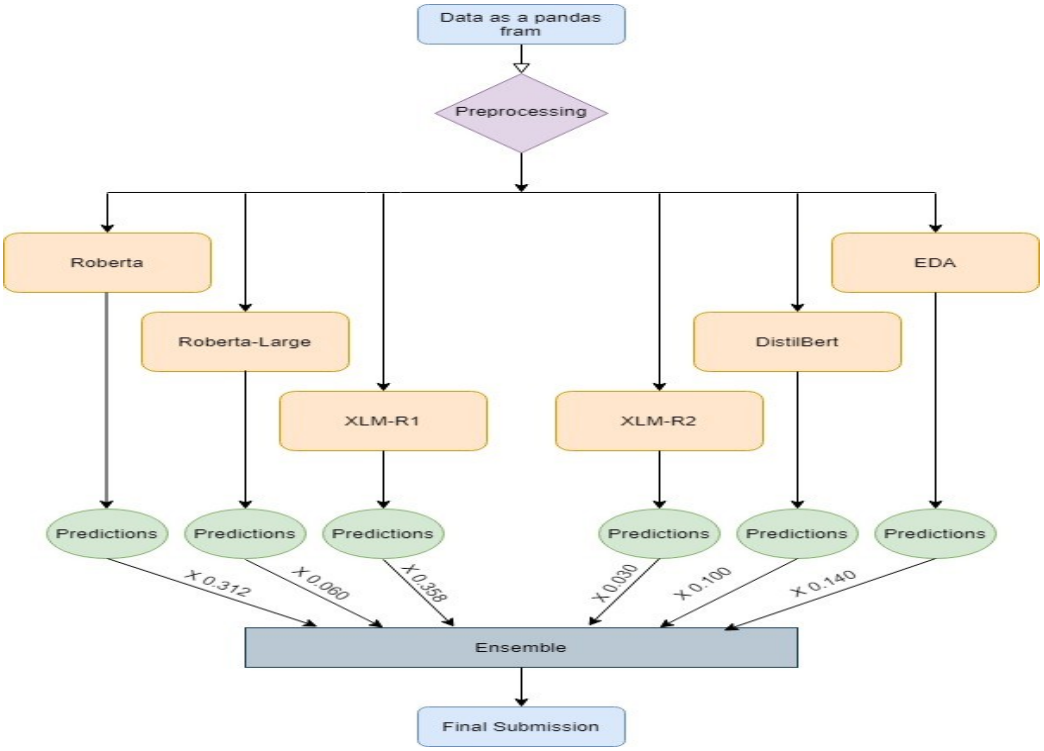
General steps followed:
1. Installing required libraries
2. Loading and Preprocessing the Datasets
3. Loading the pre-processed Model
4. Training the model on training set for some epochs
5. Predictions and Submission

Models that gave highest results: Distilbert-base-multilingual-cased XLM-RoBERTa (XLM-R). XLM-R outperforms mBERT on cross-lingual classification by up to 23% accuracy on low-resource languages. To solve the current problem we used an ensemble consisting of Roberta Large, XLM-Roberta, EDA, Distilbert-Multilingual-Cased.

Dataset

The primary data for the competition in each provided file was in the 'comment_text' column. This contains the text of a comment which has been classified as toxic or non-toxic. The training set's comments are entirely in English and come from

Civil Comments or Wikipedia talk page edits. The test data's comment_text columns are composed of multiple languages and not limited to English. The comments were also classified into following categories - Toxic, Severe toxic, Obscene, Threat, Insult, Identity hate.



Evaluation measures

Submissions are evaluated on “area under the ROC curve” between the predicted probability and the observed target.

Software and Hardware Requirements

1. Pytorch
2. Fastai framework (optional)
3. Numpy, Pandas Python libraries
4. GPU/TPU support (Google Colab/Kaggle)