Quora Insincere Question Detection

**Introduction**

Quora is an online question-answer discussion /blog platform. The motivation of this model is to identify the insincere questions (i.e.) the questions that are racially abusive, misappropriate etc. on the Quora platform before they are posted publicly.

**Problem**

The language model will be different across the regions and require specific data preparation for embedding layer. After adjusting the embedding layer, specific pattern of words needs to be learnt by the model to interpret the context. The data has class imbalance problem.

**Data Pre-Processing**

Data Source: Kaggle competition on Quora

It involves correcting several spellings on the data, abbreviating the short term to match the embedding layer and replacing most similar words to the dictionary word which are not available. Replacing numbers with hash so that they can be vectorized properly. Tokenizing each sentence into word and taking most frequent word for classification feature.

**Architecture Used**

LSTM cell with a combination of CNN with weighted F1 score as metric.

**Result**

Achieved 66.39% f1 score on the test data provided in the competition.