Final Project: Unintended Bias in Toxicity Classification

Detect toxicity across a diverse range of conversations

Problem Description:

(originally a challenge from Kaggle) Traditional toxicity models in conversational AI predicted a high likelihood of toxicity for comments contain certain identities, e.g., gay, even when those comments are not actually toxic, e.g.,I am a gay woman.

Tools:

Tensorflow/Keras, numpy, pandas, sklearn, glove, matplotlib.pyplot, seaborn

Data:

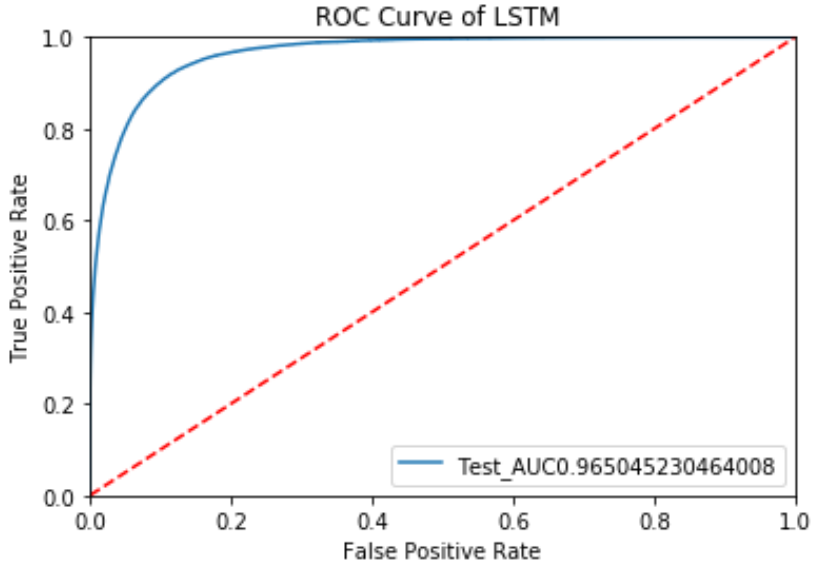
We use the dataset of Civil Comments downloaded from Kaggle. The dataset contains "comments" and main label "target" for toxicity. It also has several additional toxicity subtype attributes, which are 'severe_toxicity', 'obscene', 'identity_attack', 'insult', 'threat'.

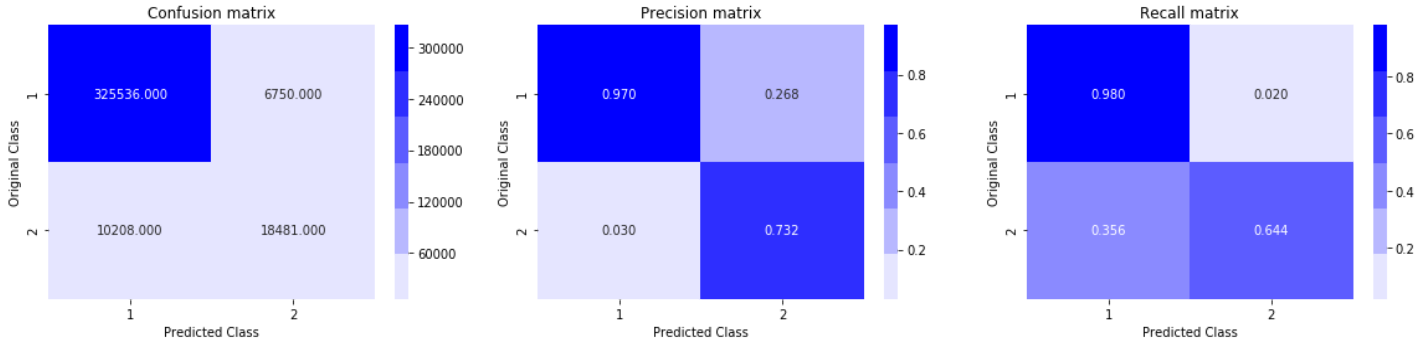
Methods:

In this project, we build a simple LSTM model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities. We train the LSTM model by using 80% of dataset, validate the model by 20% of dataset, and then predict the toxicity on some example comments. This simple LSTM model gives a prediction accuracy of 0.95 and AUC of 0.96 on validation dataset, which are relatively good performance considering its simple structure. The toxicity model can distinguish between comments with unintended bias and toxic comments with identity attack, i.e., intended bias.

Model: pre-trained Glove embedding model as encoder and two bidirectional LSTM layers as decoder.

Result:





How to run:

1. Create a folder, name it with “Colab_SMDL”
2. Create sub-folders:
    - data: download from https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
    - emb:folder to put documents of glove embeddings
3. Upload the folder “Colab_SMDL” to your drive (upload to the root)
4. Run “Project_toxicity_classificarion.ipynb”