Toxic Comments Classification (Identity Bias)

Surbhi Prasad
Karishma Chauhan
Jun 30, 2022

AGENDA

1. Situation and Data Source

Overview of goal and data

2. EDA and Preprocessing

Distribution of words in toxic comments and core distt

3. Evaluation Metric

Metric getting minimized including bias component

3. Model Comparison

Models tried and compared to pick best

Goal

Problem:
For non-toxic comments, model predicts as toxic for certain sensitive categories. Models predicted a high likelihood of toxicity for comments containing identities (e.g., “gay”), even when those comments were not actually toxic (such as “I am a gay woman”).

sentence	"seen as toxic"
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

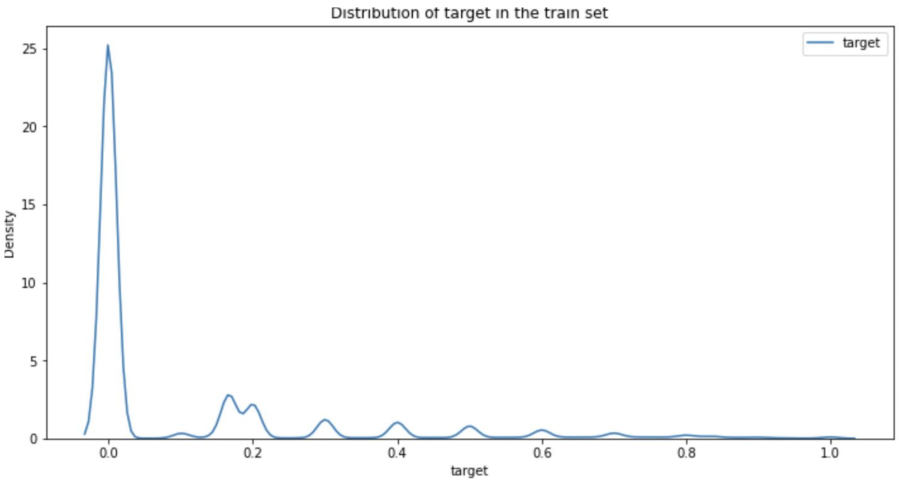
Goal
Build a model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities.

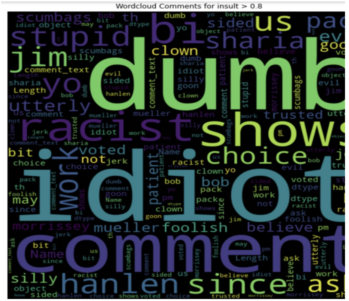
Relevant identities: *male, female, homosexual_gay_or_lesbian, Christian, Jewish, Muslim, black, white, psychiatric_or_mental_illness.*

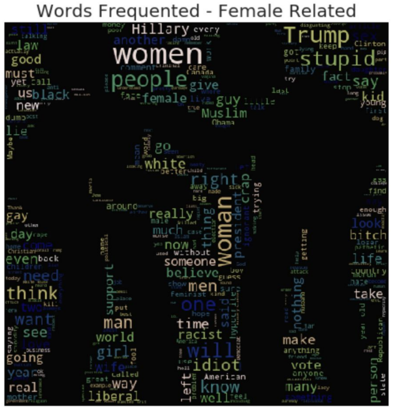
Data
At the end of 2017, the Civil Comments platform shut down and chose to make their ~2m public comments from their platform available in a lasting open archive

Exploratory Data Analysis

There are almost 70 % of data has target values<=0.1







92 % of data belong to the non-toxic class and 7 % of data belong to the toxic class

In all subgroups, there are 77.55 % of comments have NAN values.

Text Comments PreProcessing

- Normalized comments text as follows:
  - Changed capital letters to lower letters.
  - Made a table to handle different language characters.
  - De-emojized all comment texts
  - Removed extra spaces and punctuation.
  - Mapped hidden abuse words covered by ** with original words to train better

Baseline Model : TFIDF + SGD

Binary Classification Problem where we have to classify a given comment as toxic or Non-toxic.

TFIDF with SGD Classifier for different alphas were tried, Best alpha=0.0001
Penalty L2, log loss
TFID parameters:
Ngram range: (1,2), Min df: 3, smoothing

```
For values of alpha =  1e-05 The auc score on CV is: 0.810816
For values of alpha =  0.0001 The auc score on CV is: 0.8254805333333333
For values of alpha =  0.001 The auc score on CV is: 0.8034773333333332
For values of alpha =  0.01 The auc score on CV is: 0.7876821333333334
For values of alpha =  0.1 The auc score on CV is: 0.7841066666666666
For values of alpha =  1 The auc score on CV is: 0.7866538666666668
For values of alpha =  10 The auc score on CV is: 0.7896832
```

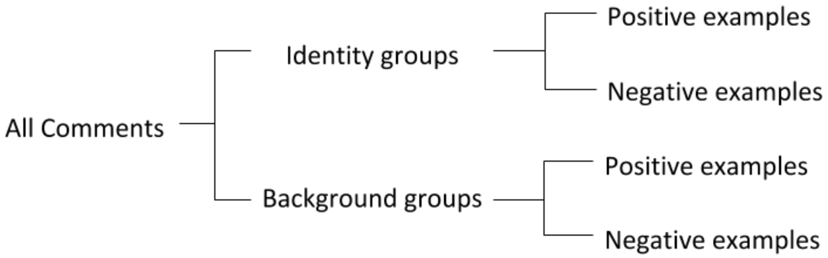
Evaluation Metric

$$\text{Score} = w_0 \, AUC_{overall} + \sum_{a=1}^{A} w_a M_p(m_{s,a})$$

where, A = number of submetrics (3)

$m_{s,a}$ = bias metric for identity subgroup s using submetric a

w_a = a weighting for the relative importance of each submetric (w values set to 0.25)



a. Subgroup AUC — This calculates AUC on only the examples from the subgroup. It represents model understanding and performance within the group itself

b. BNSP AUC — This calculates AUC on the positive examples from the background and the negative examples from the subgroup.

c. BPSN AUC — This calculates AUC on the negative examples from the background and the positive examples from the subgroup.

Power mean of all three bias metrics

Deep Learning Model: LSTM + GRU

- Text Padding to 200 words
- KFold=3
- Transformed sentence to seq of words
- Glove embeddings for Vocab, max =500000
- Batch Size=512
- Epochs=2
- Embedding vocab size: 404791

- Log Loss Function
- Adam Optimizer

Layers

1. Bidirectional LSTM
2. Linear
3. GRU
4. GRU +LSTM
5. Output Linear

Hidden Layer size= 64, 32, layers=2, dropout=0.2

```
Epoch 1: Train loss: 0.4442, BIAS AUC: 0.9021, Valid loss: 0.4323, BIAS AUC: 0.9214
Epoch 2: Train loss: 0.4280, BIAS AUC: 0.9229, Valid loss: 0.4275, BIAS AUC: 0.9231
```

```
Epoch 1: Train loss: 0.4451, BIAS AUC: 0.9016, Valid loss: 0.4311, BIAS AUC: 0.9179
Epoch 2: Train loss: 0.4285, BIAS AUC: 0.9225, Valid loss: 0.4273, BIAS AUC: 0.9237
```

```
Epoch 1: Train loss: 0.4450, BIAS AUC: 0.9026, Valid loss: 0.4319, BIAS AUC: 0.9190
Epoch 2: Train loss: 0.4282, BIAS AUC: 0.9224, Valid loss: 0.4283, BIAS AUC: 0.9207
```

Performance across identities

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
2	homosexual_gay_or_lesbian	735	0.870923	0.845379	0.971873
6	black	759	0.872285	0.867139	0.965666
7	white	1389	0.892541	0.865245	0.971977
5	muslim	814	0.917186	0.908285	0.964685
0	male	2560	0.937628	0.938210	0.962053
1	female	3501	0.940882	0.950304	0.955416
8	psychiatric_or_mental_illness	315	0.953398	0.946439	0.962670
3	christian	1896	0.958670	0.948718	0.966918
4	jewish	277	0.959537	0.935782	0.970816

What went well

1. By carefully pre-processing the data we were able to reduce the bias loss and got test auc of 0.92.

2. We were able to achieve the goal which means sensitive categories were fairly classified.

3. Improved weighted loss function as per subgroups.

What didn’t went well

1. Pre-processing takes so much time even with GPU

2. Training also takes a lot of time and had to restart notebook when memory was full.

3. Some libraries had version issues so had to change modeling so many times.

4. TFIDF Model didn’t do well for classification.

Future scope

1. Text Augmentation to include other resources
2. Ensemble of BERT and LSTM Models
3. Methods like improved Bucket Sequencing to fasten training

THANKS