MITIGATING BIAS ON DIFFERENT TOXIC COMMENT CLASSIFICATION MODELS

Armanda Lewis, Lakshmi Menon, Tamar Novetsky, Sameen Reza

DS-GA 1011 Final Project | Fall 2020

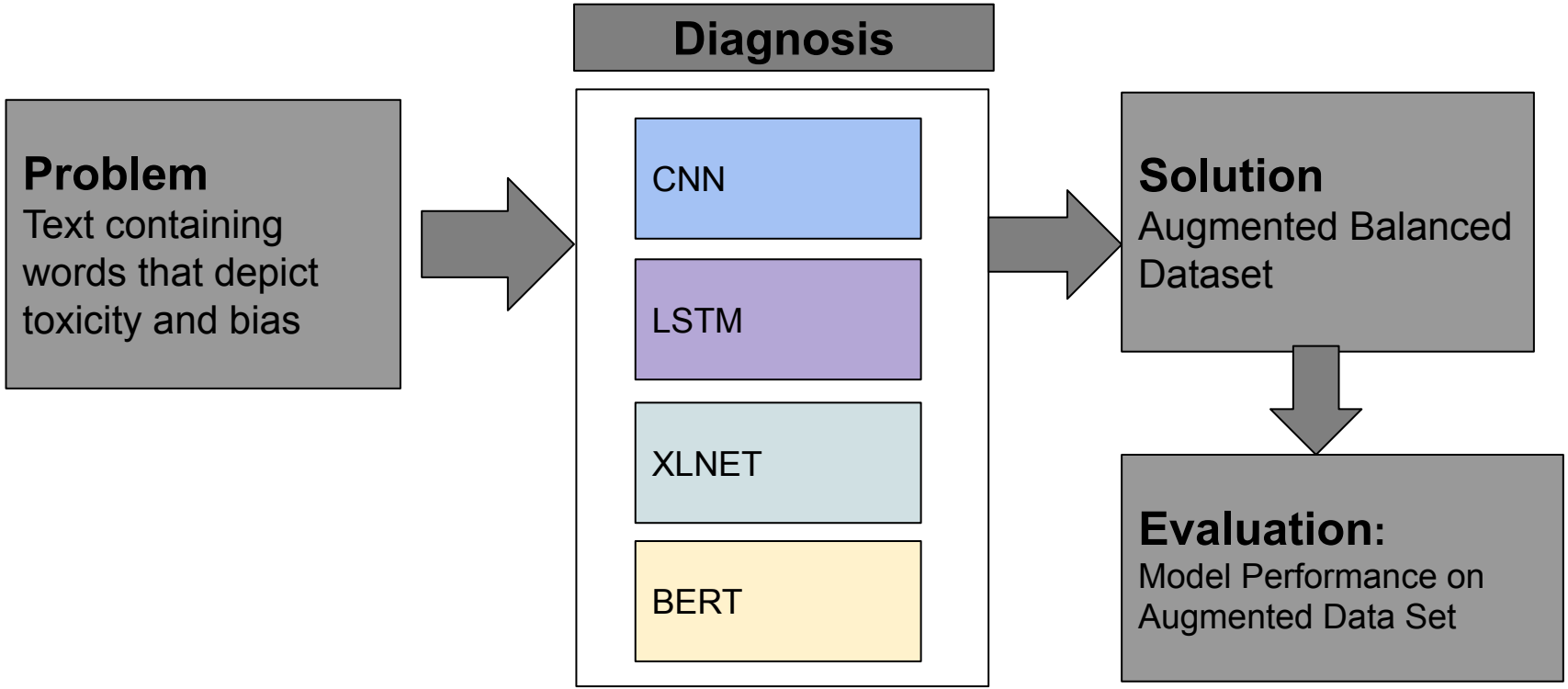
Intro and Motivation

- Bias in ML models can come from two sources:
  - Dataset
  - Models themselves
- All models have bias - that’s how they make decisions - but the problem arises when that bias comes from cultural influence or preconceived notions towards specific genders, races or religion, which violates fairness of systems and leads to harmful behavior
- Our dataset: Civil Comments
- The problem: predict toxicity while mitigating bias from the dataset - class imbalance for certain identity categories

Learned Bias

- Zhao et al (2019) found evidence of bias learned by ELMo word embeddings, which were used to create a state-of-the-art conference system which inherited the bias learned by the word embeddings.
- The ConversationAI team also found that many models incorrectly learned to associate certain identities with toxicity due to the toxic contexts in which they often occurred (Jigsaw 2018)
- Certain non-toxic comments are predicted to be toxic with baseline models, possibly due to the presence of identity which terms which tend to occur in more toxic contexts
    - "Antisemitism was prevalent in Eastern Europe once upon a time (Poland, Ukraine, Russia, etc.). Heck. The Catholic Church absolved the Jews of deicide only in 1964."
    - "Mason, Black Ernie, is just black."
    - "White people can rap too. Rap music is a genre, not a race.

Problem, Diagnosis & Solution



Bias Mitigation

- We detected bias in our original model based on different AUCs within different subgroups
- Hypothesis: This is due to differences in the distribution of different identities among the toxic comments
- Solution: data augmentation

```
Number of toxic examples in first 50K rows: 3303 (6.61%)
black, first 50K rows: 309 total, 72 toxic (0.62% of rows, 2.18% of toxic)
christian, first 50K rows: 512 total, 71 toxic (1.02% of rows, 2.15% of toxic)
female, first 50K rows: 1286 total, 169 toxic (2.57% of rows, 5.12% of toxic)
homosexual_gay_or_lesbian, first 50K rows: 356 total, 77 toxic (0.71% of rows,
2.33% of toxic)
jewish, first 50K rows: 96 total, 14 toxic (0.19% of rows, 0.42% of toxic)
male, first 50K rows: 990 total, 113 toxic (1.98% of rows, 3.42% of toxic)
muslim, first 50K rows: 150 total, 37 toxic (0.30% of rows, 1.12% of toxic)
psychiatric_or_mental_illness, first 50K rows: 171 total, 35 toxic (0.34% of
rows, 1.06% of toxic)
white, first 50K rows: 515 total, 121 toxic (1.03% of rows, 3.66% of toxic)
```

	subgroup	subgroup_auc	bpsn_auc	bnsp_auc
2	homosexual_gay_or_lesbian	0.387218	0.743217	0.549073
0	male	0.722500	0.798864	0.838488
7	white	0.750583	0.805236	0.863668
5	muslim	0.771429	0.805996	0.864494
3	christian	0.776423	0.836442	0.837509
6	black	0.796296	0.796354	0.914832
8	psychiatric_or_mental_illness	0.807692	0.869822	0.827877
1	female	0.816959	0.793329	0.889108
4	jewish	1.000000	0.884302	0.968167

Data Augmentation

- Certain groups are overrepresented in the toxic class, resulting in comments of these classes being predicted to be toxic more often
- To account for this, we include enough non-toxic samples of each subgroup so that their distribution among toxic comments matches the prior distribution on the overall dataset
- According to the formula below, we add comments iteratively for each identity until their proportion is balanced

$$n_{add} = \frac{p_{toxic} \cdot n_{total} - n_{identity}}{1 - p_{toxic}}$$

Where

n_{add} is the number of non-toxic samples of a given identity to be added

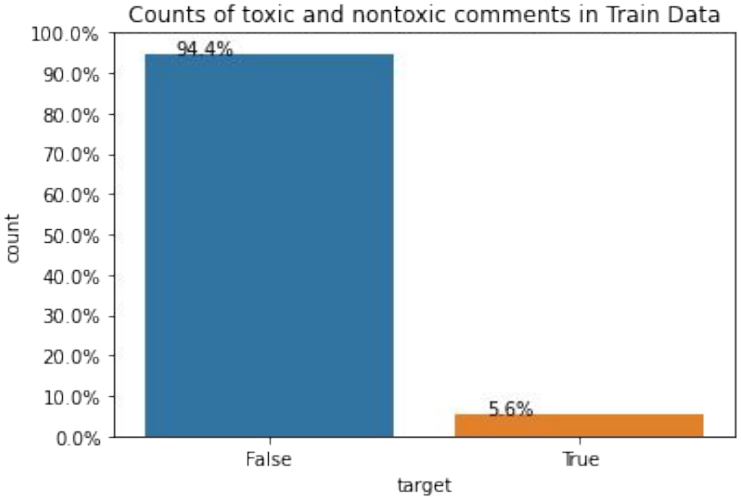
n_{total} is the total number of comments

$p_{toxic} = \frac{n_{toxic\_identity}}{n_{toxic\_total}}$ is the proportion of the given identity in toxic comments

Data Augmentation ~Method 1

Identity	% of toxic	% of total (before)	% of total (after)
asian	0.12	0.09	0.12
black	2.18	0.62	2.18
christian	2.15	1.02	2.15
female	5.12	2.57	5.12
homosexual_gay_or_lesbian	2.33	0.71	2.33
jewish	0.42	0.19	0.42
male	3.42	1.98	3.42
muslim	1.12	0.30	1.12
psychiatric_or_mental_illness	1.06	0.34	1.06
white	3.66	1.03	3.66

Initial Dataset Size: 50,000
Final Dataset Size: 58,677
Proportion of toxic comments: 5.63%



Distribution of labels before augmentation

Data Augmentation ~Method 2

Identity	% of toxic	% of total (before)	% of total (after)
asian	0.58	0.53	0.58
black	11.40	7.22	11.40
christian	7.76	6.75	7.82
female	19.36	16.32	19.67
homosexual_gay_or_lesbian	5.94	4.97	6.05
jewish	1.90	1.46	2.14
male	15.14	12.53	17.01
muslim	5.74	3.66	6.12
psychiatric_or_mental_illness	2.58	2.16	2.59
white	15.62	10.26	17.23

Initial Dataset Size: 10,000
Final Dataset Size: 16,476
Proportion of toxic comments: 30.35%

Model Performance (AUC Scores)

MODEL	TRAIN		TEST	
	Before	After	Before	After
LSTM (no pretrain)	0.933	0.951	0.911	0.919
LSTM	0.931	0.932	0.939	0.937
BERT	0.90	0.99	0.675	0.95
XLNet	0.87	0.74	0.86	0.8

Future Work

- Train models on larger datasets, as well as data from different sources
- Try data augmentation with data different from the training source
- Look at impact on intersectional categories / interactions between groups and models