Bias-Free Hate Speech Detection

Team 07:
Sanjoy Chowdhury
Satish Kumar
Konigari Rachna
Vivek Anand

Aim

To propose a novel idea for bias free hate speech detection.

Idea

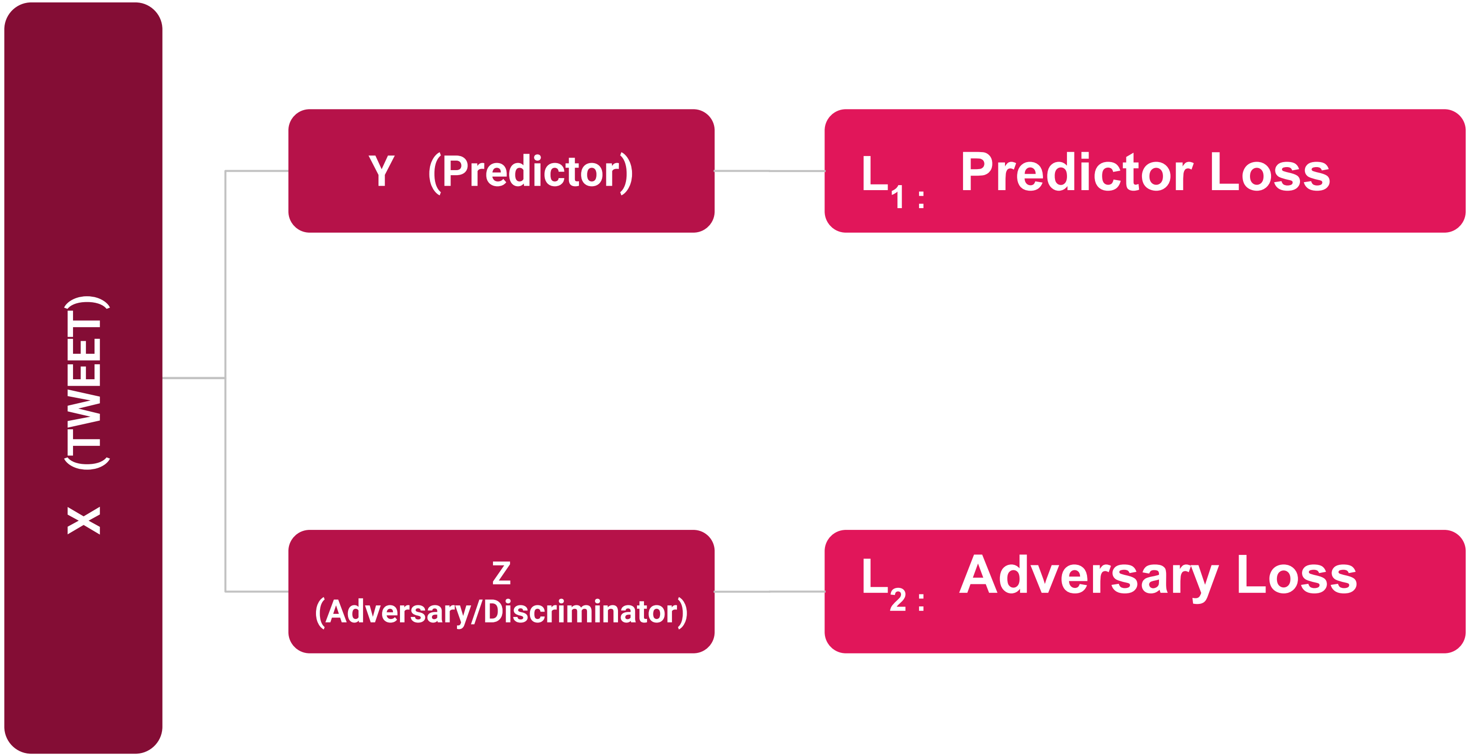
An adversarial training procedure to remove information about the sensitive attribute from the representation learnt by the neural network.

Mathematical Representation

A supervised deep learning task in which it is required to predict an output variable Y given an input variable X, while remaining unbiased with respect to some variable Z.

Here X is a given statement/corpus, Y represents whether statement is Hate or not. Z represents the set of protective features from which we want Y to be unbiased.

Diagramatic illustration



Our aim is to decrease L1 and increase L2, thereby enhancing predictor's ability to detect hate and at the same time reducing discriminator's ability to detect the protected features.

Preprocessing Dataset

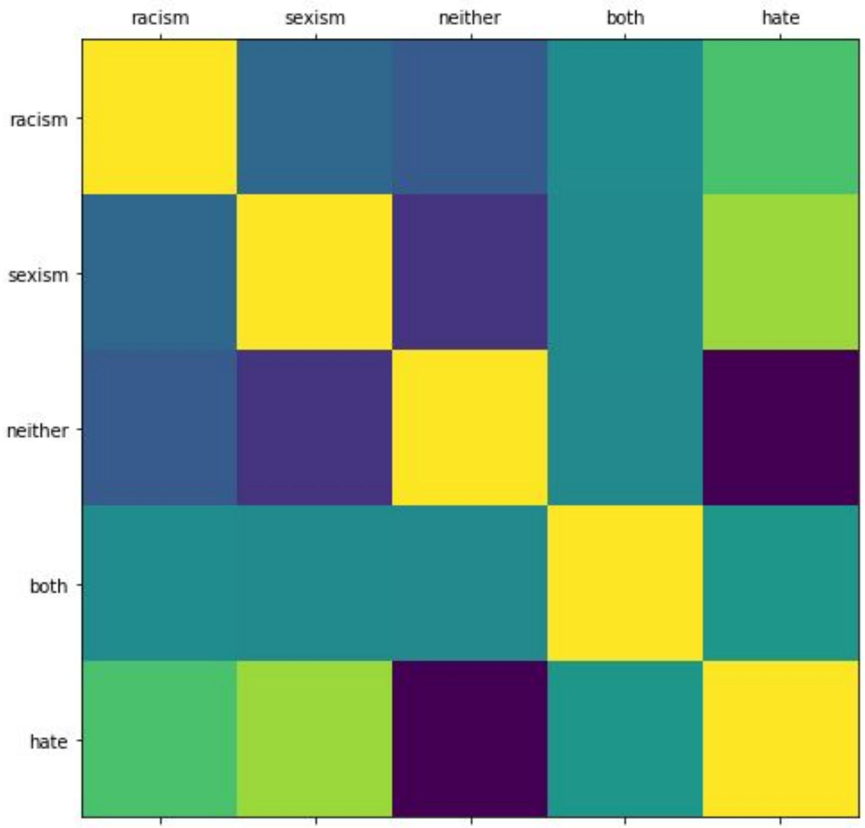
- **Labelling of dataset:** Given tagged dataset had four labels - Sexist, Racist, Neither and Both.
- We clustered sexist, racist and both labelled tweets as hate and neither as non-hate .
- Our main Objective is to predict if a tweet is Hate/Non-Hate while removing the protected feature of race.

Dataset Distribution and Structure

```
neither    5850
sexism     4341
racism     2074
both         50
Name: label, dtype: int64

hate(1) and non-hate(0) label count:-

1    6465
0    5850
Name: hate, dtype: int64

racism(1) and non-racism(0) label count:-

0    10241
1     2074
Name: racism, dtype: int64
```

Correlation Matrix of our Dataset

As the shade of the colour in the correlation matrix gets **brighter**, it indicates the **stronger correlation** between the two labels.



Dataset Distribution and Structure

	tweet	label	racism	sexism	neither	both	hate
0	http t.co zxbzv39jru cat perform trick one min...	neither	0	0	1	0	0
1	catarybertt lorexplo feminazi detect	sexism	0	1	0	0	1
2	thequinnspiraci hey poke organ think stream th...	neither	0	0	1	0	0
3	watch shia militia beat peshmerga death though...	racism	1	0	0	0	1
4	nicolesantucci wipe smug smile face kat rudebi...	sexism	0	1	0	0	1
5	gemmanoon thing open interpret numer boolean d...	neither	0	0	1	0	0
6	sschink teh_maxh guess anyways.	neither	0	0	1	0	0
7	raikonl finalev mja333 heheheh. liter wrangl b...	neither	0	0	1	0	0
8	i'm ponder get leo friend. look shelter pic ad...	neither	0	0	1	0	0
9	main cours go vacuous narcissist like mkr	sexism	0	1	0	0	1
10	damn cyber saudi-adjac nation jaxblast sorri s...	sexism	0	1	0	0	1
11	mkr sexist four six team fourth instant restau...	neither	0	0	1	0	0
12	trigger_check rather entertain see get threat ...	neither	0	0	1	0	0
13	untouchablesh femin lobbi equal veto actually....	sexism	0	1	0	0	1
14	silvermillsi freebsdgirl everi day find harder...	neither	0	0	1	0	0
15	cavusseyit euklidproof kobane_ypg certifi inma...	racism	1	0	0	0	1
16	stop sass ogbaisteid mkr	neither	0	0	1	0	0
17	know that interest idea. mayb get promot tweet...	neither	0	0	1	0	0

Feature Space

$F_y : \{F_{y1}, F_{y2}, F_{y3}, \ldots\ldots F_{yk}\}$ and $F_z : \{F_{z1}, F_{z2}, F_{z3}, \ldots\ldots F_{zk}\}$

There are two possibilities:

1. $F_y \cap F_z = \emptyset$

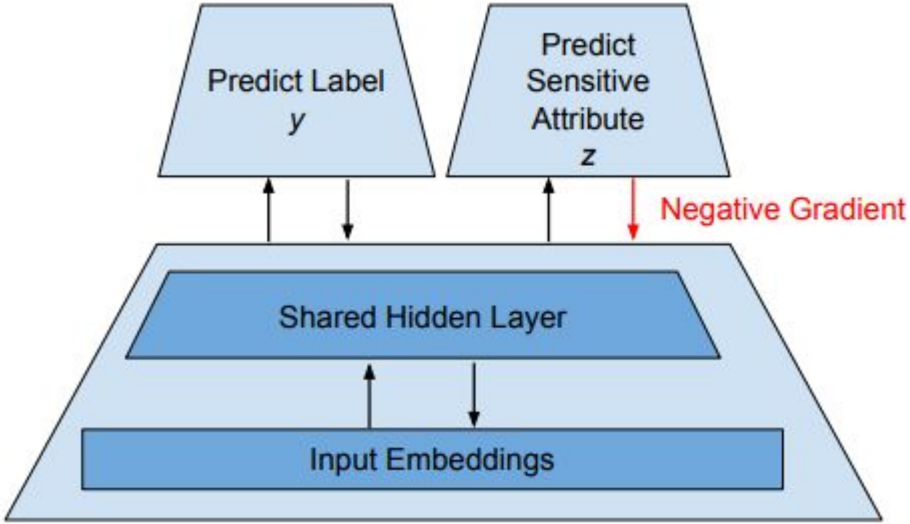
   This implies that predictor and adversary don't share any common feature and hence removing adversarial features won't affect predictor's ability.

2. $F_y \cap F_z \neq \emptyset$

   When there is an intersecting set of features between the discriminator and the adversary, normally removing adversarial features will affect the predictor.

   But if the network is able to learn some new features that helps in prediction, then predictor's ability isn't affected.

Network Architecture of Approach 1



Y = <Hate, Non-hate>

Z = <Set of protective features>

(Eg. Sexist/Non-Sexist,Racist/Non-Racist)

Loss Function

We need to minimize

$$\mathbf{L_1 - \lambda * L_2}$$

Where

- $\boldsymbol{\lambda}$ is a constant weight given to $\mathbf{L_2}$
- $\mathbf{L_1}$ is the loss incurred by the model which predicts hate
- $\mathbf{L_2}$ is the loss incurred by the model which predicts bias

Explanation of Approach 1

We have converted the problem to a **multi-loss optimization** problem where predictor's loss i.e., L1, has to be minimized and at the same time adversarial loss i.e., L2 has to be maximized. Total loss is represented as **L**

There are many mathematical approaches to

1. $\mathbf{L = L_1 - \lambda * L_2}$
2. $\mathbf{L = L_1 + \lambda * (1/L_2)}$

The network will try to minimize the loss **L**, which is the linear combination of the respective losses.

Predictor’s Result of Approach 1

Hate Without Bias	Precision	Recall	F1 Score	support
0	0.92	**0.88**	0.90	1166
1	0.89	**0.93**	0.91	1297

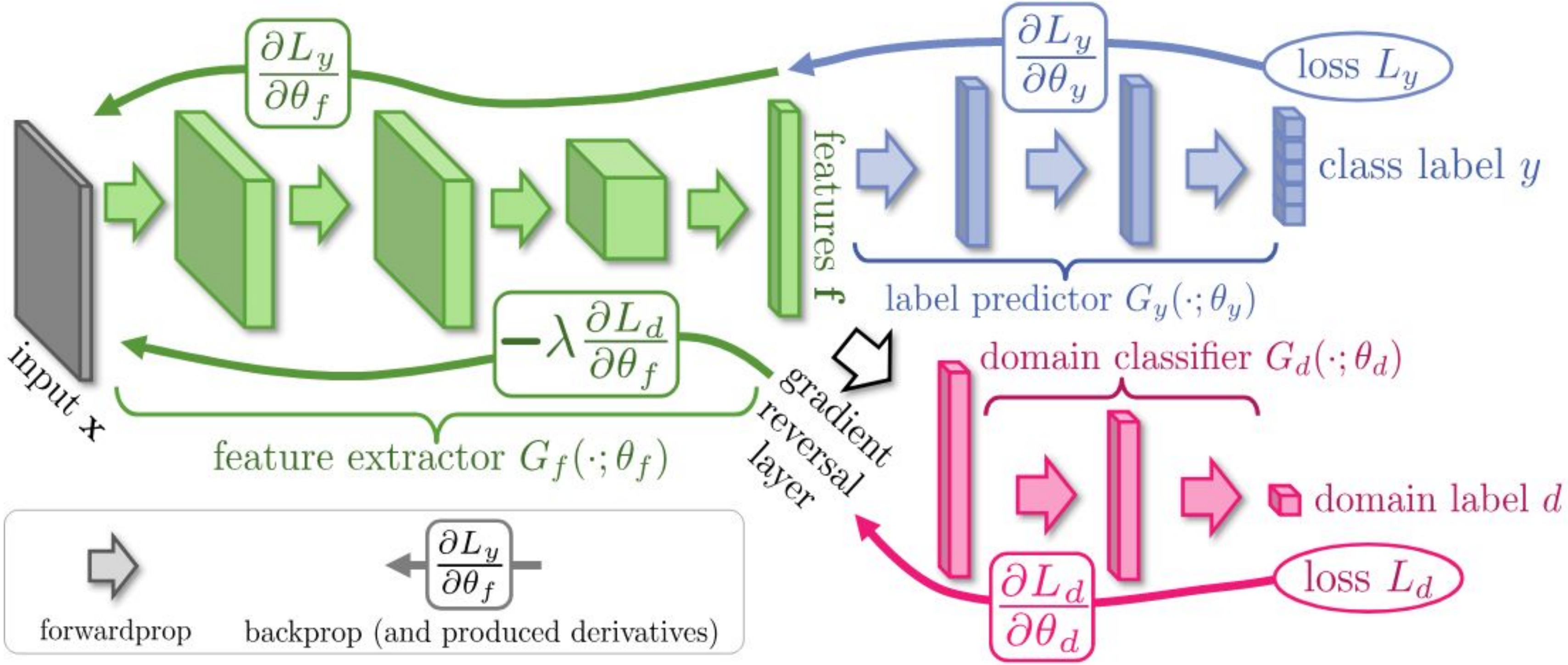
Hate with Bias	Precision	Recall	F1 Score	support
0	0.86	**0.93**	0.89	1166
1	0.93	**0.86**	0.90	1297

Bias Result of Approach 1

Racism in Hate With Bias Model	Precision	Recall	F1 Score	support
0	0.91	**0.95**	0.93	2033
1	0.70	**0.76**	0.62	430

Racism in Hate without Bias Model	Precision	Recall	F1 Score	support
0	1.00	**0.00**	0.00	2032
1	0.18	**1.00**	0.30	431

Network Architecture of Approach 2



Network Architecture of Approach 2

- Network starts learning some common features for predictor and adversary and later as seen in figure, then it splits into two separate classifiers , namely predictor and adversary.

- While backpropagating, predictor’s loss will be passed as it is, but adversary’s loss is back propagated through the gradient reversal layer.

- This helps the network to remove the protected features.

Predictor’s Result of Approach 2

Hate Without Bias	Precision	Recall	F1 Score	support
0	0.92	**0.88**	0.90	1166
1	0.89	**0.93**	0.91	1297

Hate with Bias	Precision	Recall	F1 Score	support
0	0.84	**0.92**	0.87	1147
1	0.92	**0.84**	0.88	1316

Bias Result of Approach 2

Racism in Hate With Bias Model	Precision	Recall	F1 Score	support
0	0.91	**0.95**	0.93	2033
1	0.70	**0.76**	0.72	430

Racism in Hate without Bias Model	Precision	Recall	F1 Score	support
0	0.00	**0.00**	0.00	2053
1	0.17	**1.00**	0.29	410

Papers Referred

- Alex Beutel, Jilin Chen, Zhe Zhao, Ed H. Chi - Data decisions and Theoretical Implications when adversarially learning fair representations.
- Brian Hu Zhang, Blake Lemoine, Margaret Mitchell - Mitigating Unwanted biases with adversarial learning.