Detection of gender bias in Hate Speech Classifier

Presented By:

Group 11
Archana Kumari - 1906011
Priyanshu Raj - 1906012
Garge Archana Atul - 1906124

Under Guidance Of :
Dr Jyoti Prakash Singh

29 April, 2022

Fairness And Bias

- Fairness: Absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics.
- Bias: When scientific or technological decisions are based on a narrow set of systemic, structural or social concept and norms, the resulting technology can privilege certain groups and harm others

Objective

- Identification of gender bias in textual hate speech
- Preprocessing of data to metigate gender bias
- Improve fairness in data

Problems We Are Addressing

- Gender bias in hate speech
- Pre-trained embeddings exaggerates bias

Related works

Table 1: Some of the potential work

Authors	aim	Method Used	models
[Park et al., 2018] Park, Ji Ho and Shin, Jamin and Fung, Pascale	Reducing Gender Bias in Abusive Language Detection	Debiased Word Embeddings (DE), Gender Swap (GS), Bias fine-tuning (FT)	CNN, GRU, alpha-GRU
[Bolukbasi et al., 2016] Bolukbasi, Tolga and Chang, Kai-Wei and Zou, James Y and Saligrama, Venkatesh and Kalai, Adam T	De-bias Gender Stereotype in Word-Embedding	Gender Subspace, cosine similarity, w2vNEWS embedding	Hard De-Biasing (neutralize and equalize), Soft Bias Correction

Our Approach

Data Gathering

- manual labeling of hate speech dataset in three classes
  0 for male
  1 for female
  2 for Neutral

Analysis using

- Machine Learning Algorithms
- Deep Learning Algorithms

Dataset before labeling

	created_at	id	id_str	text	truncated	entities/hashtags/0/text	entities/hashtags/0/indices/0	entitie
0	Mon Mar 02 10:22:18 +0000 2015	5.72341E+17	5.723410e+17	Sassy...? More like femme bots than killer blo...	FALSE	mkr	54	
1	Mon Mar 02 10:36:01 +0000 2015	5.72345E+17	5.723450e+17	I've had better looking shits than these two! ...	FALSE	MKR2015	46	
2	Mon Mar 02 10:14:29 +0000 2015	5.72339E+17	5.723390e+17	The girls can cook for me anytime. Just not ho...	FALSE	MKR	70	
3	Sun Mar 01 02:05:37 +0000 2015	5.71854E+17	5.718540e+17	The face of very ugly promo girls ! Faces like...	FALSE	mkr	61	
4	Mon Mar 02 10:39:21 +0000 2015	5.72345E+17	5.723450e+17	@mykitchenrules Elegant and beautiful?Cheap an...	FALSE	mkr	126	

5 rows × 661 columns

Figure 1: Initial Dataset

Dataset after labeling

	id	text	Garge Archana Atul	Archana Kumari	Priyanshu Raj	Final
0	1	Sassy...? More like femme bots than killer blo...	1	1.0	1	1.0
1	2	I've had better looking shits than these two! ...	1	1.0	2	1.0
2	3	The girls can cook for me anytime. Just not ho...	1	1.0	2	1.0
3	4	The face of very ugly promo girls ! Faces like...	1	1.0	1	1.0
4	5	@mykitchenrules Elegant and beautiful?Cheap an...	1	1.0	1	1.0

Figure 2: Final Dataset

New Data-set

Table 2: The detailed of data samples for each of the classes.

Text Categories	**No. of Text per class**
Target to Male	83
Target to Female	1539
Not target any gender	1041
Total No. of Data	2663

Reports of ML Algorithms

ML Algorithms	class	precision	recall	f1-score
SVM (linear)	0	0	0	0
	1	0.85	0.83	0.84
	2	0.73	0.80	0.76
SVM (poly)	0	0	0	0
	1	0.79	0.86	0.83
	2	0.74	0.70	0.72
SVM (rbf)	0	0	0	0
	1	0.83	0.80	0.83
	2	0.71	0.83	0.76
SVM (rbf)	0	0	0	0
	1	0.83	0.80	0.83
	2	0.71	0.83	0.76

Reports of ML Algorithms

ML Algorithms	class	precision	recall	f1-score
Logistic Regression	0	0	0	0
	1	0.83	0.84	0.83
	2	0.73	0.77	0.75
Ada Boosting	0	0.20	0.09	0.13
	1	0.86	0.69	0.77
	2	0.63	0.83	0.72
XGB Boosting	0	0.67	0.09	0.16
	1	0.86	0.80	0.83
	2	0.71	0.83	0.76
Gradient Boosting	0	0.22	0.09	0.13
	1	0.86	0.80	0.83
	2	0.72	0.82	0.77

Reports of Deep Learning Models

Models	class	precision	recall	f1-score
Simple RNN	0	0.07	0.06	0.06
	1	0.74	0.82	0.78
	2	0.66	0.57	0.61
GRU	0	0	0	0
	1	0.81	0.81	0.81
	2	0.68	0.74	0.71
CNN	0	0	0	0
	1	0.83	0.78	0.80
	2	0.67	0.78	0.72

Definitions Of Fairness

- Statistical Parity
- Overall Accuracy Equality
- Conditional Procedure Accuracy

Fairness Calculation for Deep learning Models

Models	Statistic Parity
Simple RNN	0.07
GRU	0.197
CNN	0.349

Overall Accuracy for Deep Learning models

the proportion of cases misclassified difference

Models	Difference
Simple RNN	0.561
GRU	0.100
CNN	0.035

Conditional Procedure Accuracy for Deep Learning models

in terms of True Positive Rate and False Positive Rate

Models	**TP diff**	**FP diff**
Simple RNN	0.972	0
GRU	0.828	0.649
CNN	0.794	0

Method we used for debiasing

- modifying the actual vector-representations of words
- plotting words in dataset vector distance from 'male' and 'female' word
- minimising distance of words from 'male' and 'female' word

Methodology

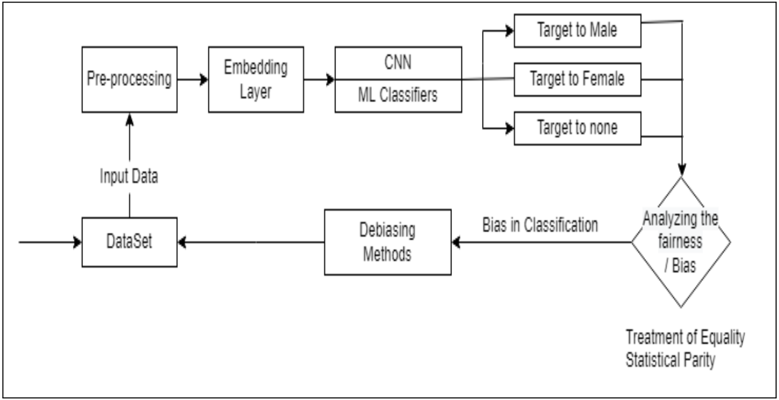

Figure 3: Workflow

Plot of all words

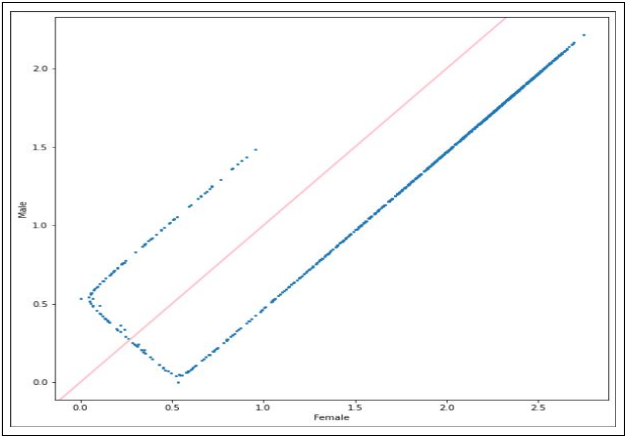

Figure 4: all words plotted wrt. 'male' and 'female' word

Plot of all words

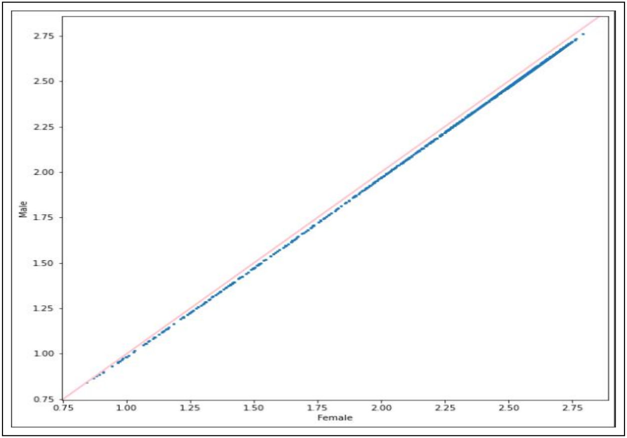

Figure 5: after shifting all words on x equal to y axis

Results After shifting

Models	class	precision	recall	f1-score
Simple RNN	0	0.08	0.06	0.07
	1	0.70	0.81	0.75
	2	0.64	0.50	0.56
GRU	0	0	0	0
	1	0.77	0.77	0.77
	2	0.65	0.70	0.67
CNN	0	0	0	0
	1	0.58	1	0.73
	2	0	0	0

Future Scope

- Refinement of modification vector for better fairness score
- POS minimise the dependence of adjectives and verbs on subjects

References

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Thank You