Easy Adaptation to Mitigate Gender Bias in Multilingual Text Classification

Xiaolei Huang

University of Memphis

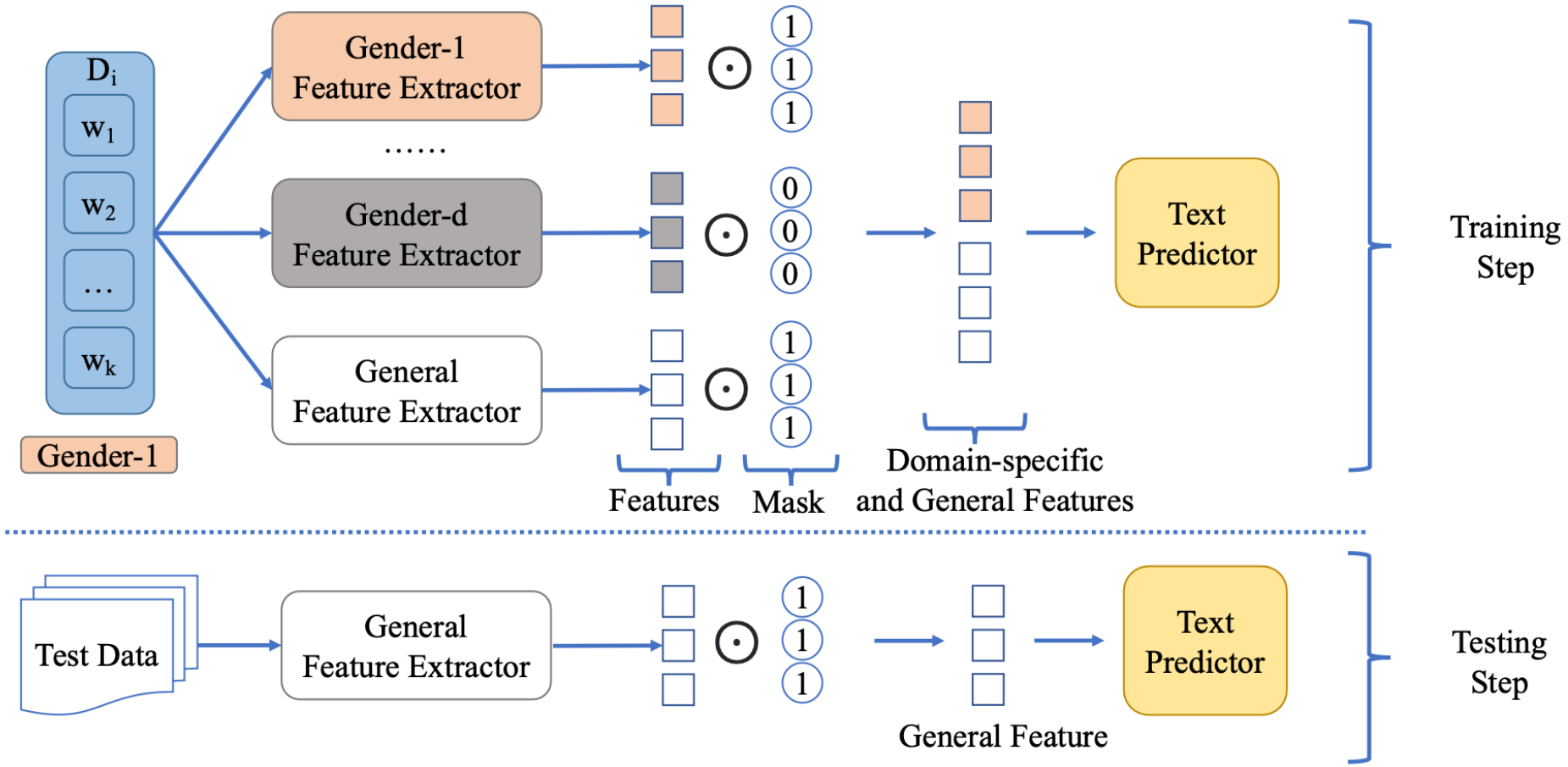
Multilingual Fairness in Text Classifiers

- *Multilingual*: Extensive methods have proposed new studies to reduce biases in text classifier, while limited studies have been evaluated under multilingual settings.

- *Group Fairness*[1]: document classifiers are defined as biased if the classifiers perform better for documents of some groups than for documents of other groups.

1. Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

Easy Adaptation Framework



Easy Adaptation Framework

Feature Augmentation[1]	Domain-independent encoder: General. Domain-dependent encoder: Female, Male.	
Training	Domain mask	Female: $[F_{general}, F_{female}, 0]$ Male: $[F_{general}, 0, F_{male}]$
Testing	General-only	$[F_{general}, 0, 0]$

1. Daumé III, H. (2007, June). Frustratingly Easy Domain Adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (pp. 256-263).

Multilingual Datasets of Different Tasks

- Hate Speech[1]
  - hate (or related) or not.
- Rating Prediction[2]
  - how likely a customer would purchase a product: positive / negative.

Source	Lang	Docs	Tokens	F-Ratio	L-Ratio
HS	EN	44,253	20.533	.498	.355
	IT	2,361	19.848	.310	.235
	PT	1,852	20.007	.554	.222
	ES	4,831	20.660	.455	.357
Review	EN	358,219	48.553	.398	.930
	FR	324,358	37.102	.429	.931
	DE	115,367	38.224	.430	.928
	DA	882,080	49.829	.475	.886

1. Huang, X., Xing, L., Dernoncourt, F., & Paul, M. (2020, May). Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. LREC.
2. Hovy, D., Johannsen, A., & Søgaard, A. (2015, May). User review sites as a resource for large-scale sociolinguistic studies. In Proceedings of the 24th international conference on World Wide Web (pp. 452-461).

Evaluation

Regular	1) F1-macro 2) AUC
Fair	**Sum** (*FPED* + *FNED*) of equality differences (ED) of false positive/negative rates (*FP/FN-ED*). $FPED = \sum_{g \in \mathrm{G}} \lvert FPR_d - FPR \rvert$,where G is the gender and d is a gender group (e.g., female).

Evaluation

Performance on the HS and Review data in percentage. A lower fair score is better. The Delta-R and -F are improvements over the regular (-R) and fair (-F) baselines respectively. Negative Delta scores over the fair indicate percentage of mitigating biases, and lower scores means more bias mitigation.

Review (%)		English			French			German			Danish		
Methods		F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair
B-Reg	LR	87.1	98.3	4.2	85.1	97.9	7.7	86.1	97.9	7.6	88.5	98.4	6.2
	RNN	87.1	97.6	5.2	80.6	95.3	1.5	80.4	95.5	3.7	86.8	94.8	3.1
	BERT	93.3	99.3	4.9	91.6	**99.1**	4.6	91.2	**98.2**	3.5	**94.0**	98.8	3.9
B-Fair	LR-Blind	87.1	98.3	3.6	85.2	97.9	7.6	86.0	97.9	5.9	90.5	98.4	**1.4**
	RNN-Blind	89.6	98.5	4.5	81.7	96.0	5.1	82.0	96.8	3.2	85.8	95.7	2.5
	BERT-Blind	93.4	99.3	4.3	**91.7**	**99.1**	3.9	89.5	98.5	3.6	93.0	99.1	1.9
	RNN-IW	87.9	98.8	**2.8**	81.6	97.2	4.4	84.5	97.6	3.2	86.2	97.6	1.8
	RNN-Adv	88.0	98.1	5.2	83.4	96.9	4.9	85.4	97.4	3.0	88.7	97.6	1.9
Ours	LR-DA	87.3	98.4	**2.8**	85.3	97.9	1.7	85.1	98.0	4.3	87.6	98.5	3.3
	RNN-DA	89.2	98.6	4.1	83.1	95.9	**0.9**	81.2	96.6	3.5	89.2	98.2	1.9
	BERT-DA	**93.6**	**99.4**	3.3	**91.7**	99.0	3.4	**91.4**	97.8	**2.7**	93.7	**99.2**	1.7
Delta-R (%)		1.0	.4	-28.7	1.1	0.2	-56.5	0	0.3	-29.1	0.5	1.3	-47.7
Delta-F (%)		.9	0.2	-16.7	2.3	0.2	-61.4	0.5	-0.2	-7.4	1.5	1.0	21.1

Hate Speech (%)		English			Spanish			Italian			Portuguese		
Methods		F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair
B-Regular	LR	81.5	89.3	6.2	66.6	80.9	27.2	54.8	75.5	21.1	65.3	75.2	12.8
	RNN	82.0	89.0	5.4	65.3	70.0	25.9	62.3	70.7	30.9	60.8	75.9	44.1
	BERT	84.3	**92.0**	4.9	65.9	73.8	15.6	57.1	70.3	12.9	70.1	79.6	19.9
B-Fair	LR-Blind	81.5	89.1	5.4	67.3	**81.0**	25.9	54.8	75.5	20.7	62.2	73.9	9.6
	RNN-Blind	82.8	89.8	5.1	64.9	63.8	14.2	56.4	**76.4**	22.9	62.2	74.9	20.6
	BERT-Blind	84.0	91.9	3.7	65.5	72.8	14.9	57.2	71.2	23.2	72.4	**81.8**	26.4
	RNN-IW	83.8	98.4	3.8	54.0	58.9	13.4	64.1	74.7	21.9	63.8	74.7	30.7
	RNN-Adv	82.9	90.6	4.1	54.6	64.8	12.0	57.9	70.9	22.1	69.8	75.8	23.1
Ours	LR-DA	81.0	88.6	4.3	71.5	79.7	18.5	62.9	71.1	17.8	67.4	79.0	11.8
	RNN-DA	82.1	89.1	4.7	66.5	70.9	22.8	62.8	72.3	25.6	68.8	77.1	11.7
	BERT-DA	**84.4**	91.4	**2.2**	**73.8**	78.3	**10.1**	**67.2**	74.9	**12.4**	**74.8**	78.3	**9.0**
Delta-R (%)		-0.1	-0.4	-32.1	7.1	1.9	-25.2	10.7	0.8	-14.0	7.5	1.6	-57.7
Delta-F (%)		-0.6	-2.5	-15.5	15.2	11.8	6.6	10.7	-1.3	-16.1	6.4	2.5	-50.9

Takeaways

- Fair-aware methods reduce gender biases under multilingual settings, while not generally improve accuracy.
- Translating gender-sensitive tokens from English can be effective for multilingual settings.
- Adaptation method is Easy and Effective.

- https://github.com/xiaoleihuang/DomainFairness