Easy Adaptation to Mitigate Gender Bias in Multilingual Text Classification

Xiaolei Huang
University of Memphis

1. Fair-aware methods reduce gender biases under multilingual settings, while not generally improve accuracy.

2. Translating gender-sensitive tokens from English can be effective for multilingual settings.

3. Adaptation method is Easy and Effective.

https://github.com/xiaoleihuang/DomainFairness

Fairness / Bias

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.

Multilingual

First study: adaptation method and multilingual evaluation for fair-aware classifiers.

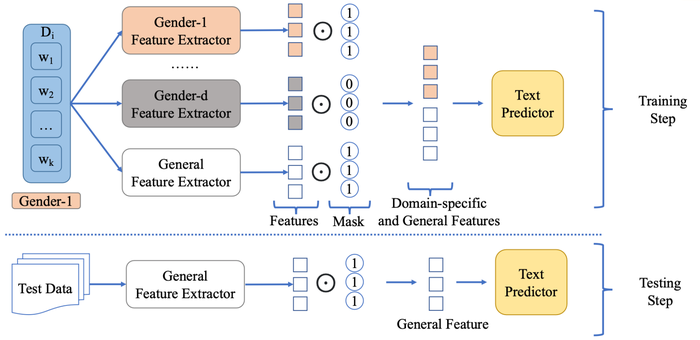
Two tasks

Hate Speech[2]

RecSys[3]

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

Framework



Feature Augmentation[4]	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]

Eval & Apps

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

scores : improvements of our method over regular (1st row) & fair-aware (2nd row) baselines.

Hate Speech

English			Spanish			Italian			Portuguese		
F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair
-0.1	-0.4	-32.1	7.1	1.9	-25.2	10.7	0.8	-14.0	7.5	1.6	-57.7
-0.6	-2.5	-15.5	15.2	11.8	6.6	10.7	-1.3	-16.1	6.4	2.5	-50.9

RecSys

English			French			German			Danish		
F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair	F1-macro	AUC	Fair
1.0	.4	-28.7	1.1	0.2	-56.5	0	0.3	-29.1	0.5	1.3	-47.7
.9	0.2	-16.7	2.3	0.2	-61.4	0.5	-0.2	-7.4	1.5	1.0	21.1

Reference:

1. Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.
2. Huang, X., Xing, L., Dernoncourt, F., & Paul, M. (2020, May). Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. LREC.
3. 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).
4. 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).