VLDB2023
vancouver

Through the Fairness Lens: Experimental Analysis and Evaluation of Entity Matching



Nima Shahbazi, Nikola Danevski, Abolfazl Asudeh,
Fatemeh Nargesian, Divesh Srivastava


Agenda

- Motivation
- Part I: Fairness Evaluation in the context of EM
- Part II: Evaluation
  1. Entity-matching Approaches, datasets
  2. A few (representative) Evaluation Results
- Part III: Lessons and Discussions

(Direct) Social applications of EM

Airline Security

E-commerce

Health care

Things can go wrong! (esp. for marginalized people)

No-fly List

Airline Security

High-value customers list

E-commerce

How about EM for Union & Join?

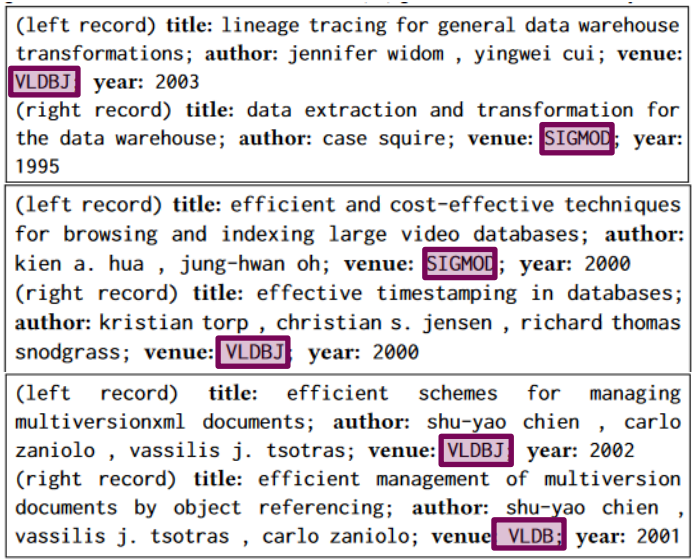
Part I: (Group) Fairness Evaluation in the Context of EM

Fairness Measures for EM Evaluation

- Sensitive attributes (e.g., race, gender)
- Demographic Groups (e.g., female)
  - Non- intersectional (e.g., male) v.s. Intersectional groups (e.g. Black male)
  - Non-overlapping (e.g., gender) v.s. Overlapping groups (e.g., Research domain (DB, ML))

Single v.s. Pairwise Fairness

- **Pairwise nature of EM** (a differentiation from ML classification)
  - Each row in the test set: a record $c = (e_i, e_j, h, y)$
  - Single Fairness: fairness is evaluated for a group **g**
    - a record (in test-set) is “counted” for group **g** if at least one of the pairs belong to **g**
  - Pairwise Fairness: fairness is evaluated for a pair **(g,g’)** group.



Name	Description	Equation ($\forall g_i \in \mathcal{G}$)
Accuracy Parity (AP)	requires the independence of matchers's accuracy from groups	$Pr(h(e,e') = y \mid g_i) \simeq Pr(h(e,e') = y)$
Statistical Parity (SP)	requires the independence of the matcher from groups	$Pr(h(e,e') = \text{`}M\text{'} \mid g_i) \simeq Pr(h(e,e') = \text{`}M\text{'})$
[1]True Positive Rate Parity (TPRP)	a.k.a *Equal Opportunity*; in the group of true matches requires the independence of match predictions from groups	$Pr(h(e,e') = \text{`}M\text{'} \mid g_i, y = \text{`}M\text{'}) \simeq Pr(h(e,e') = \text{`}M\text{'} \mid y = \text{`}M\text{'})$
False Positive Rate Parity (FPRP)	in the group of true non-matches, requires the independence of match predictions from groups	$Pr(h(e,e') = \text{`}M\text{'} \mid g_i, y = \text{`}N\text{'}) \simeq Pr(h(e,e') = \text{`}M\text{'} \mid y = \text{`}N\text{'})$
[1]False Negative Rate Parity (FNRP)	in the group of true matches, requires the independence of non-match predictions from groups	$Pr(h(e,e') = \text{`}N\text{'} \mid g_i, y = \text{`}M\text{'}) \simeq Pr(h(e,e') = \text{`}N\text{'} \mid y = \text{`}M\text{'})$
True Negative Rate Parity (TNRP)	in the group of true non-matches, requires the independence of non-match predictions from groups	$Pr(h(e,e') = \text{`}N\text{'} \mid g_i, y = \text{`}N\text{'}) \simeq Pr(h(e,e') = \text{`}N\text{'} \mid y = \text{`}N\text{'})$
[1]Equalized Odds (EO)	in both groups of true matches and true non-matches requires the independence of match predictions from groups	$Pr(h(e,e') = \text{`}M\text{'} \mid g_i, y = \text{`}M\text{'}) \simeq Pr(h(e,e') = \text{`}M\text{'} \mid y = \text{`}M\text{'})$ $Pr(h(e,e') = \text{`}M\text{'} \mid g_i, y = \text{`}N\text{'}) \simeq Pr(h(e,e') = \text{`}M\text{'} \mid y = \text{`}N\text{'})$
[1]Positive Predictive Value Parity (PPVP)	among the pairs predicted as match requires the independence of true matches from groups	$Pr(y = \text{`}M\text{'} \mid h(e,e') = \text{`}M\text{'}, g_i) \simeq Pr(y = \text{`}M\text{'} \mid h(e,e') = \text{`}M\text{'})$
[1]Negative Predictive Value Parity (NPVP)	among the pairs predicted as non-match, requires the independence of true non-matches from groups	$Pr(y = \text{`}N\text{'} \mid h(e,e') = \text{`}N\text{'}, g_i) \simeq Pr(y = \text{`}N\text{'} \mid h(e,e') = \text{`}N\text{'})$
[1]False Discovery Rate Parity (FDRP)	among the pairs predicted as match, requires the independence of true non-matches from groups	$Pr(y = \text{`}N\text{'} \mid g_i, h(e,e') = \text{`}M\text{'}) \simeq Pr(y = \text{`}N\text{'} \mid h(e,e') = \text{`}M\text{'})$
[1]False Omission Rate Parity (FORP)	among the pairs predicted as non-match, requires the independence of true matches from groups	$Pr(y = \text{`}M\text{'} \mid g_i, h(e,e') = \text{`}N\text{'}) \simeq Pr(y = \text{`}M\text{'} \mid h(e,e') = \text{`}N\text{'})$

Name	Description	Equation ($\forall g_i \in \mathcal{G}$)
Accuracy Parity (AP)	requires the independence of matchers's accuracy from groups	$Pr(h(e,e')=y \mid g_i) \simeq Pr(h(e,e')=y)$
Statistical Parity (SP)	requires the independence of the matcher from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i) \simeq Pr(h(e,e')=\text{`}M\text{'})$
[1]**True Positive Rate Parity (TPRP)**	a.k.a *Equal Opportunity*; in the group of true matches requires the independence of match predictions from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}M\text{'})$
False Positive Rate Parity (FPRP)	in the group of true non-matches, requires the independence of match predictions from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}N\text{'})$
[1]**False Negative Rate Parity (FNRP)**	in the group of true matches, requires the independence of non-match predictions from groups	$Pr(h(e,e')=\text{`}N\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}N\text{'} \mid y=\text{`}M\text{'})$
True Negative Rate Parity (TNRP)	in the group of true non-matches, requires the independence of non-match predictions from groups	$Pr(h(e,e')=\text{`}N\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}N\text{'} \mid y=\text{`}N\text{'})$
[1]**Equalized Odds (EO)**	in both groups of true matches and true non-matches requires the independence of match predictions from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}M\text{'})$ $Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}N\text{'})$
[1]**Positive Predictive Value Parity (PPVP)**	among the pairs predicted as match requires the independence of true matches from groups	$Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}M\text{'}, g_i) \simeq Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}M\text{'})$
[1]**Negative Predictive Value Parity (NPVP)**	among the pairs predicted as non-match, requires the independence of true non-matches from groups	$Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}N\text{'}, g_i) \simeq Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}N\text{'})$
[1]**False Discovery Rate Parity (FDRP)**	among the pairs predicted as match, requires the independence of true non-matches from groups	$Pr(y=\text{`}N\text{'} \mid g_i, h(e,e')=\text{`}M\text{'}) \simeq Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}M\text{'})$
[1]**False Omission Rate Parity (FORP)**	among the pairs predicted as non-match, requires the independence of true matches from groups	$Pr(y=\text{`}M\text{'} \mid g_i, h(e,e')=\text{`}N\text{'}) \simeq Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}N\text{'})$

Name	Description	Equation ($\forall g_i \in \mathcal{G}$)
Accuracy Parity (AP)	requires the independence of matchers's accuracy from groups	$Pr(h(e,e')=y \mid g_i) \simeq Pr(h(e,e')=y)$
Statistical Parity (SP)	requires the independence of the matcher from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i) \simeq Pr(h(e,e')=\text{`}M\text{'})$
[1]True Positive Rate Parity (TPRP)	a.k.a *Equal Opportunity*; in the group of true matches requires the independence of match predictions from groups	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}M\text{'})$
False Positive Rate Parity (FPRP)	in the group of true non-matches requires	$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}N\text{'})$
[1]False Negative Rate Parity (FNRP)		$Pr(h(e,e')=\text{`}N\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}N\text{'} \mid y=\text{`}M\text{'})$
True Negative Rate Parity (TNRP)		$Pr(h(e,e')=\text{`}N\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}N\text{'} \mid y=\text{`}N\text{'})$
[1]Equalized Odds (EO)		$Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}M\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}M\text{'})$ $Pr(h(e,e')=\text{`}M\text{'} \mid g_i, y=\text{`}N\text{'}) \simeq Pr(h(e,e')=\text{`}M\text{'} \mid y=\text{`}N\text{'})$
[1]Positive Predictive Value Parity (PPVP)	among the pairs predicted as match requires the independence of true matches from groups	$Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}M\text{'}, g_i) \simeq Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}M\text{'})$
[1]Negative Predictive Value Parity (NPVP)	among the pairs predicted as non-match, requires the independence of true non-matches from groups	$Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}N\text{'}, g_i) \simeq Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}N\text{'})$
[1]False Discovery Rate Parity (FDRP)	among the pairs predicted as match, requires the independence of true non-matches from groups	$Pr(y=\text{`}N\text{'} \mid g_i, h(e,e')=\text{`}M\text{'}) \simeq Pr(y=\text{`}N\text{'} \mid h(e,e')=\text{`}M\text{'})$
[1]False Omission Rate Parity (FORP)	among the pairs predicted as non-match, requires the independence of true matches from groups	$Pr(y=\text{`}M\text{'} \mid g_i, h(e,e')=\text{`}N\text{'}) \simeq Pr(y=\text{`}M\text{'} \mid h(e,e')=\text{`}N\text{'})$

Class Imbalance in EM: among the $O(n^2)$ pairs, (usually) only $O(n)$ of them are match

Unfairness (Disparity) -- Division

$$F_{\alpha,\beta}^{(d)}(g_i) = \max\left(0\ , 1 - \frac{Pr(\alpha \mid \beta, g_i)}{Pr(\alpha \mid \beta)}\right)$$

- E.g., True Positive Rate Parity (TPRP):

$$F = \max(0,1 - \frac{TPR(g_i)}{TPR(all)})$$

- Rule of thumb: unfairness more than 20% is not acceptable

$$F \leq 0.2$$

Part II: Evaluation

Entity Matchers

- Rule-based Matchers: 1
- Non-Neural Matchers: 7
- Neural Matchers: 5

Datasets

- Structured: 4
  - including 2 semi-synthetic social datasets
- Dirty: 2
- Textual: 2

We created two social datasets:

- **NoFlyCompas**: Used COMPAS dataset and created a no-fly list (impact of *over-representation*)
- **FacultyMatch**: Created using CSRankings faculty profiles in China and Germany (impact of *name similarities*)

NoFlyCompas

- W-B population in US ~ 75%-13%. In no-fly-list ~ 50%-50% (**Black is over-represented**)

- **Non-neural matchers were dominant**, both on overall performance and on fairness
  - Reason: better utilization of the structure of data.
    Also reported in [Mudgal2018]

	TPR		Disparity		FDR		Disparity	
Matcher	Afr.	Cauc.	sub	div	Afr.	Cauc.	sub	div
DeepMatcher	0.89	0.86	-0.03	-0.03	0.20	0.18	0.02	0.11
Ditto	0.76	0.82	0.06	0.08	0.31	0.22	0.09	0.41
GNEM	0.84	0.84	0.00	0.00	0.17	0.09	0.08	0.88
HierMatcher	0.72	0.74	0.02	0.10	0.22	0.16	0.06	0.38
MCAN	0.54	0.57	0.03	0.05	0.19	0.05	0.14	2.8

A FP result by Ditto

(left record) **firstName:** James **lastName:** Brown **race:** African-American
(right record) **firstName:** Samanthai **lastName:** Browne **race:** African-American

Faculty-Match

Matcher	TPR		Disparity		PPV		Disparity	
	cn	de	sub	div	cn	de	sub	div
DeepMatcher	0.48	0.72	**0.23**	**0.50**	0.79	0.87	0.08	0.11
Ditto	0.59	0.85	**0.26**	**0.44**	0.77	0.94	0.17	**0.22**
GNEM	0.78	0.90	0.12	0.15	0.83	0.92	0.08	0.11
HierMatcher	0.47	0.78	**0.31**	**0.66**	0.78	0.89	0.11	0.14
MCAN	0.40	0.70	**0.30**	**0.75**	0.86	0.94	0.08	0.09
DTMatcher	0.95	0.90	-0.05	-0.05	0.89	0.98	0.09	0.10
LinRegMatcher	0.33	0.23	-0.09	-0.43	0.44	0.96	**0.52**	**1.18**
LogRegMatcher	0.95	0.88	-0.07	-0.08	0.93	1.0	0.07	0.07
NbMatcher	0.99	0.99	0.00	0.00	0.03	0.58	**0.55**	**18.3**
RfMatcher	0.96	0.89	-0.06	-0.08	0.98	0.99	0.01	0.01
SvmMatcher	0.95	0.87	-0.07	-0.09	0.94	0.99	0.05	0.05

A FP example (Ditto)

(left record) **fullName:** Qingming Huang **country:** cn
(right record) **fullName:** Qing-Hu Huang **country:** cn

A FN example (Ditto)

(left record) **fullName:** LinLin Shen **country:** cn
(right record) **fullName:** Linlin phen **country:** cn

Highlights from Extended Experiments

- **Neural matchers** were more unfair on **structured** data
  - Relying on **pre-trained** language models and embeddings

(left record) **song:** Tequila Loves Me; **artist:** K. Chesney
(right record) **song:** Likes Me; **artist:** K. Chesney

  - not fully considering the dataset **structures**

(left record) **title:** efficient schemes for managing multiversionxml documents; **author:** shu-yao chien , carlo zaniolo , vassilis j. tsotras; **venue:** VLDBJ; **year:** 2002
(right record) **title:** efficient management of multiversion documents by object referencing; **author:** shu-yao chien , vassilis j. tsotras , carlo zaniolo; **venue:** VLDB; **year:** 2001

- **Neural matchers** failed on **Textual** data
- Putting a high-weight on **proxy attributes** causes unfairness

(left record) **title:** guest editorial; **author:** alon y. halevy; **venue:** VLDBJ; **year:** 2002
(right record) **title:** guest editorial ; **author:** vijay atluri , anupam joshi , yelena yesha; **venue:** VLDBJ; **year:** 2003

  - Reason: Different “behaviors” between groups; Insufficient coverage for “minorities”

Sensitivity to the choice of matching threshold

	Dataset	Non-neural						Neural				
		DtMatcher	LinRegMatcher	LogRegMatcher	NbMatcher	RfMatcher	SvmMatcher	DeepMatcher	Ditto	Gnem	HierMatcher	Mcan
TPRP	iTunes-Amazon	0	0	2.4	0	2.2	2.4	3.9	9.3	1	6.9	2.4
	Cameras	1	0	8.4	2.8	8.7	7.1	3.3	2.8	1	2.6	3.6
	Dblp-Acm	0	0	0	0	0	0	0	2	0	0	0
	Dblp-Scholar	0	0	0	0	0	1	2.4	2	0	2.2	2.4
PPVP	iTunes-Amazon	0	0	0	0	2	0	1.7	5.2	0	2	1.4
	Cameras	1	0	5.8	4.5	4.6	3.7	3.4	2.4	1.7	4.6	3.6
	Dblp-Acm	0	0	0	0	0	2.6	0	0	0	0	0
	Dblp-Scholar	0	0	1	1	1	1.4	1	1.4	0	2.4	1

Part III: Lessons and Discussions

- *Call for action* to collect EM benchmarks on societal applications

- Inherent issues in social data
  - Over/under-representation
  - Textual (name) similarity

We created two social datasets:
- **NoFlyCompas**: impact of *over-representation*
- **FacultyMatch**: impact of *name similarities*

Guide for EM Researchers and Practitioners

	Rules of Thumb
Structured datasets	– Non-neural matchers are preferred – Obtain attributes with min correlation with sensitive attributes – Minimize Representation bias in training data – Make sure the model is not putting high weights on only a few attributes
Textual& dirty datasets	– Neural matchers are preferred – Obtain additional (unbiased) features – Use unbiased pretrained models – Minimize Representation bias in training data – Considering their sensitivity, try out different matching thresholds and select the most fair/accurate one
Fairness measure: TPRP and PPVP are usually preferred (see § 3.5 and § 5.3.2)	
Use an ensemble of matchers (for single sensitive attributes with exclusive values): construct a set of matchers; for each group use the matcher with best performance on it (using separate test sets for each group)	

Thank you!

Code Repository: