Master's Thesis

Machine Learning for Disaster Event Classification from Social Media Images
—
Transnational Fairness with Development Indexes

Cem Kozcuer
2023-02

Supervisor: Prof. Dr. Felix Bießmann
Department VI — Computer Science and Media
Berliner Hochschule für Technik

Reviewer: Prof. Dr. Frank Haußer
Department II — Math - Physics - Chemistry
Berliner Hochschule für Technik

abstract

contributions

method:
fairness in ML
on transnational level
with development indexes

fairness investigation:
on the MEDIC dataset

extended dataset:
MEDIC with country data

outline

1. **context**: dataset / problem / related work
2. **method**: quantifying fairness in transnational disaster response
3. **results**
4. **evaluation**
5. **outlook**

MEDIC:

detect and assess disasters
from images in social media worldwide

application context: real-time disaster response system AIDR[1]

assist humanitarian help allocating response and resources

MEDIC dataset: Alam et al.[2], August 2021

task labels:

disaster types	—> *flood, hurricane, fire, earthquake*, etc.
informativeness	—> *informative, not informative*
humanitarian	—> *infrastructure damage, rescue volunteering, not humanitarian*, etc.
damage severity	—> *little or none, mild damage, severe damage*

[1] AIDR: Artificial Intelligence for Disaster Response, 2014 (published at *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web*)

[2] MEDIC: A Multi-Task Learning Dataset for Disaster Image Classification

1 context – **problem**

problem:

how to investigate fairness on the MEDIC dataset?

subject of fairness?

sensitive variables / socio-economic data not available (usually *gender, age, ethnicity,...*)

geo-diversity alone does not exhibit underlying problems of inequality

fairness for ML in DRM is under-explored

1 context – **problem**

problem:

how to investigate fairness on the MEDIC dataset?

subject of fairness?

sensitive variables / socio-economic data not available (usually *gender, age, ethnicity,...*)

geo-diversity alone does not exhibit underlying problems of inequality

fairness for ML in DRM is under-explored

solution:

development indexes provide socio-economic data on transnational / country level

development indexes can be proxies for or actual sensitive variables for fairness investigation

1 context – **problem**

problem:

how to investigate fairness on the MEDIC dataset?

subject of fairness?

sensitive variables / socio-economic data not available (usually *gender, age, ethnicity,...*)

geo-diversity alone does not exhibit underlying problems of inequality

fairness for ML in DRM is under-explored

solution:

development indexes provide socio-economic data on transnational / country level

development indexes can be proxies for or actual sensitive variables for fairness investigation

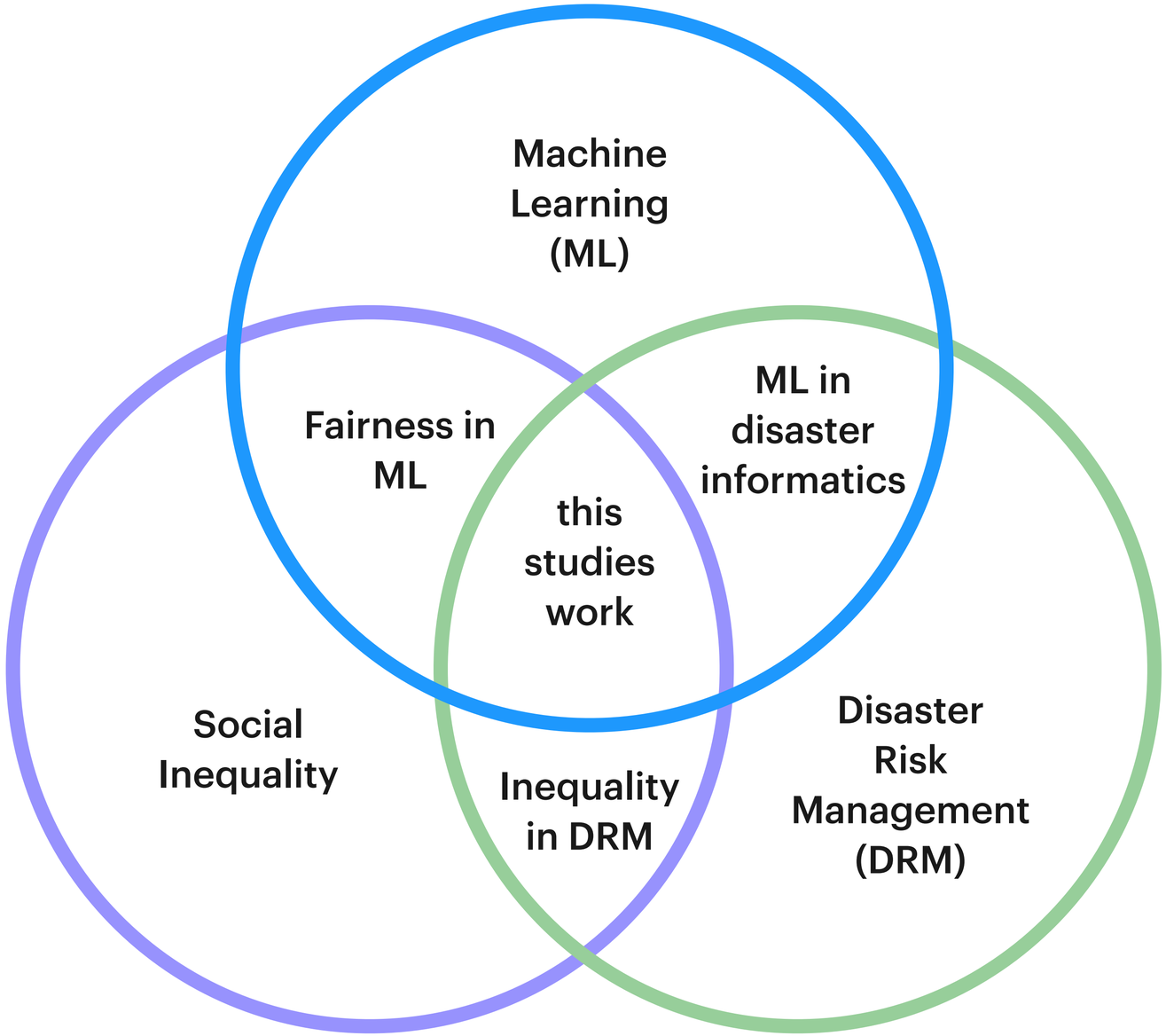
method:

group countries by development indexes and evaluating group fairness

subject of fairness are countries instead of individuals

fairness is related
to global inequalities
than to regional

1 context - related work



1 context - **related work**

study	socio-economic factors (/fairness)	social media data	DRM context	machine learning application	outside of USA	transnational level
Li et al 2013	X	X				
Zou et al 2017	X	X	X			
Wang et al 2021	X	X	X			
Dargin et al 2021	X	X	X			
Zhu et al 2021	X	X	X		X	
Shankar et al 2017	geo-diversity			X	X	continents
DeVries et al 2019	geo-diversity + income			X	X	continents
Goya et al 2022	geo-diversity + income			X	X	continents
(MEDIC) Alam et al 2021, etc.		X	X	X	X	X
This study	X	X	X	X	X	X

2 method: quantifying fairness in transnational disaster response

– development indexes as sensitive attributes

attributes and factors for successful disaster response via social media:

economical resilience

information and communications technology resilience

information and communications technology access

political freedom of media

cultural usage of media

health and welfare

2 method: quantifying fairness in transnational disaster response

– development indexes as sensitive attributes

attributes and factors for successful disaster response via social media:

economical resilience

information and communications technology resilience

information and communications technology access

political freedom of media

cultural usage of media

health and welfare

combination of development indexes as sensitive variables:

Human Development Index
(composite index of life expectancy, education and income per capita)

Democracy Index
(index of 60 questions answered by experts about pluralism, consensus about government, political participation, democratic culture and civil liberties)

ICT Development Index
(composite of 11 indicators that describe developments in information and communication technologies in order measure the global digital divide or disparities in access to internet and information technology)

2 method: quantifying fairness in transnational disaster response

– development indexes as sensitive attributes

development index	economic resilience	health and welfare	political freedom	cultural media usage	ICT resilience	ICT access
Human Development Index	X	X				
Democracy Index			X	X		
Internet and Communications Development Index					X	X

2 method: quantifying fairness in transnational disaster response

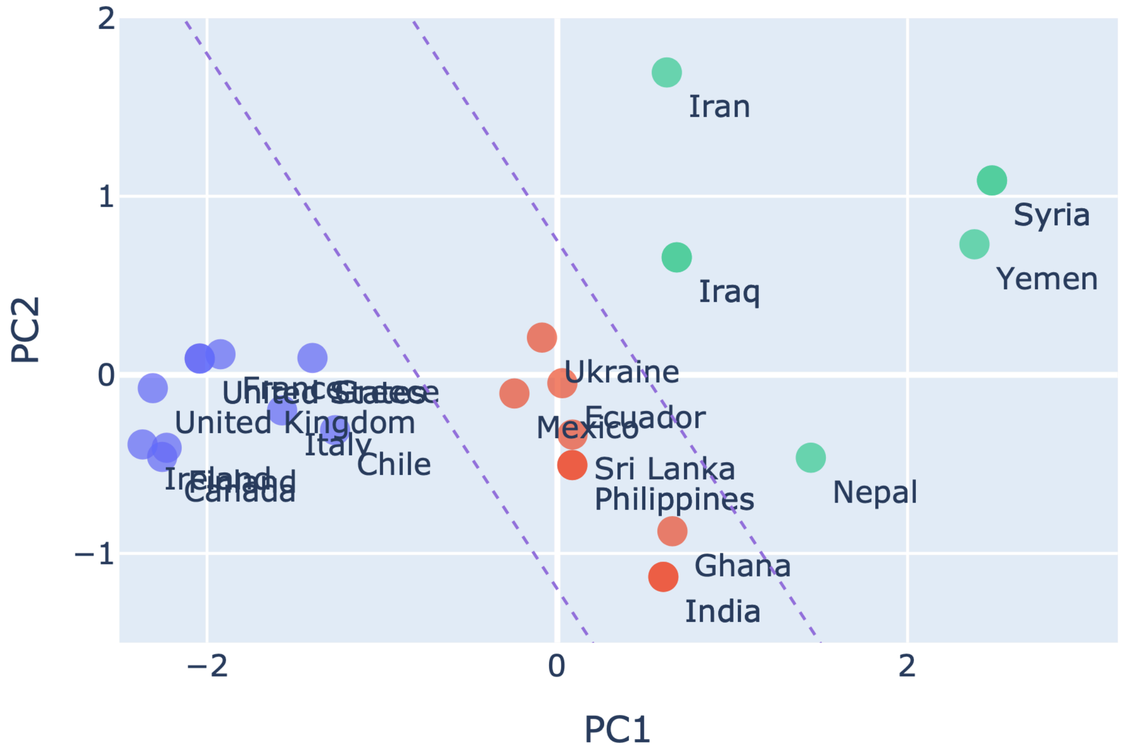
– grouping of involved countries with PCA on development indexes

97% of variance explained within first 2 PCs

Group A (blue):
Canada, United States, United Kingdom, Italy,
Greece, Ireland, Chile, France, Finland

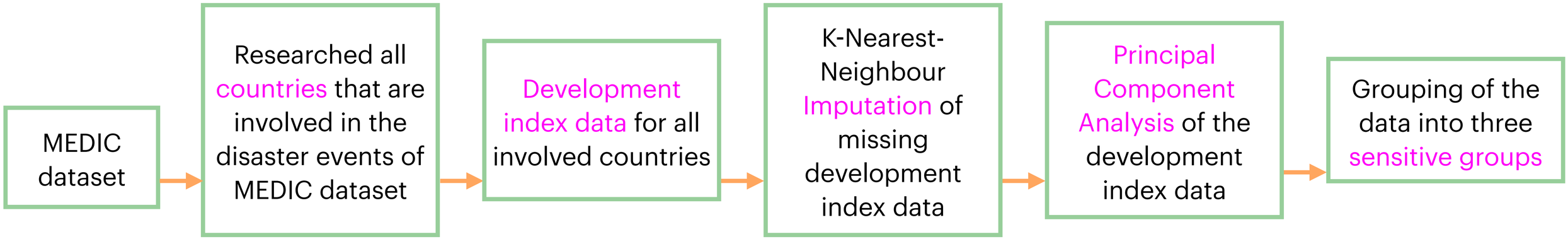
Group B (red):
India, Ecuador, Ukraine, Sri Lanka, Philippines,
Ghana, Mexico

Group C (green):
Iraq, Syria, Iran, Nepal, Yemen



2 method: quantifying fairness in transnational disaster response

– grouping of involved countries with PCA on development indexes



– fairness notions of study

selection of fairness notion:

ethical / normative problem

2 method: quantifying fairness in transnational disaster response

– fairness notions of study

selection of fairness notion:

ethical / normative problem

performance metrics of MEDIC:

(accuracy)
precision
recall
F1-score

fairness notion of this study:

based on same metrics as authors of MEDIC:

Predictive Parity <— precision
Equal Opportunity <— recall
F1-Score Parity <— F1-score

intention:

introduction of a new method

high comparability to original results

– fairness notions of study

Predictive Parity:

$$P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{A}) = P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{B})$$

2 method: quantifying fairness in transnational disaster response

– fairness notions of study

Predictive Parity:

$$P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{A}) = P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{B})$$

Equal Opportunity:

$$P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{A}) = P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{B})$$

2 method: quantifying fairness in transnational disaster response

– fairness notions of study

Predictive Parity:

$$P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{A}) = P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{B})$$

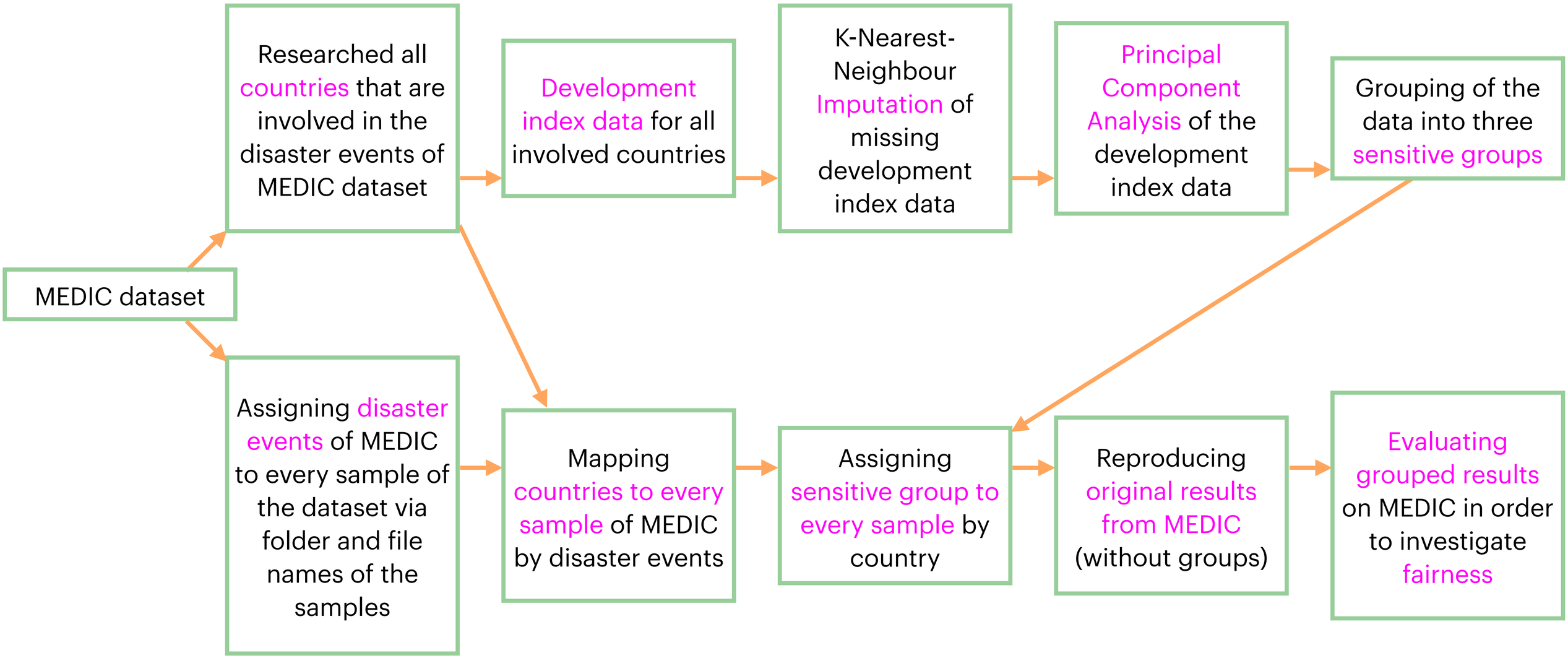
Equal Opportunity:

$$P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{A}) = P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{B})$$

F1-Score Parity:

$$2 \times \frac{P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{A}) \times P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{A})}{P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{A}) + P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{A})} = 2 \times \frac{P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{B}) \times P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{B})}{P(Y = 1 \mid \hat{Y} = 1, X \in \mathcal{B}) + P(\hat{Y} = 1 \mid Y = 1, X \in \mathcal{B})}$$

3 results – complete pipeline of study



3 results – updated MEDIC dataset for fairness

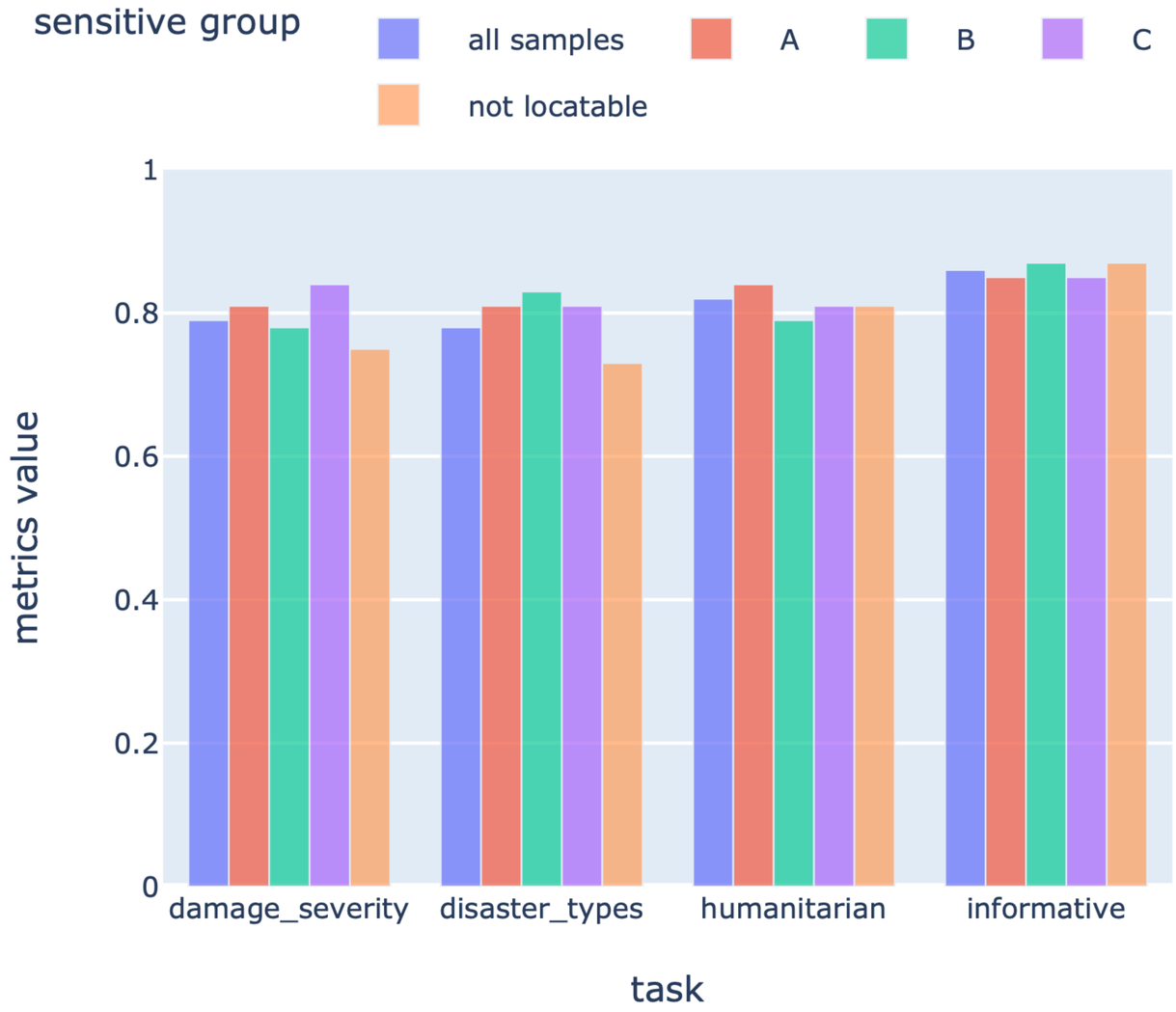
original schema

field	description
image_id	id of the sample based on the file name
event_name	name of the dataset where sample comes from
image_path	file path of the sample
disaster_types	classification task
informative	classification task
humanitarian	classification task
damage_severity	classification task

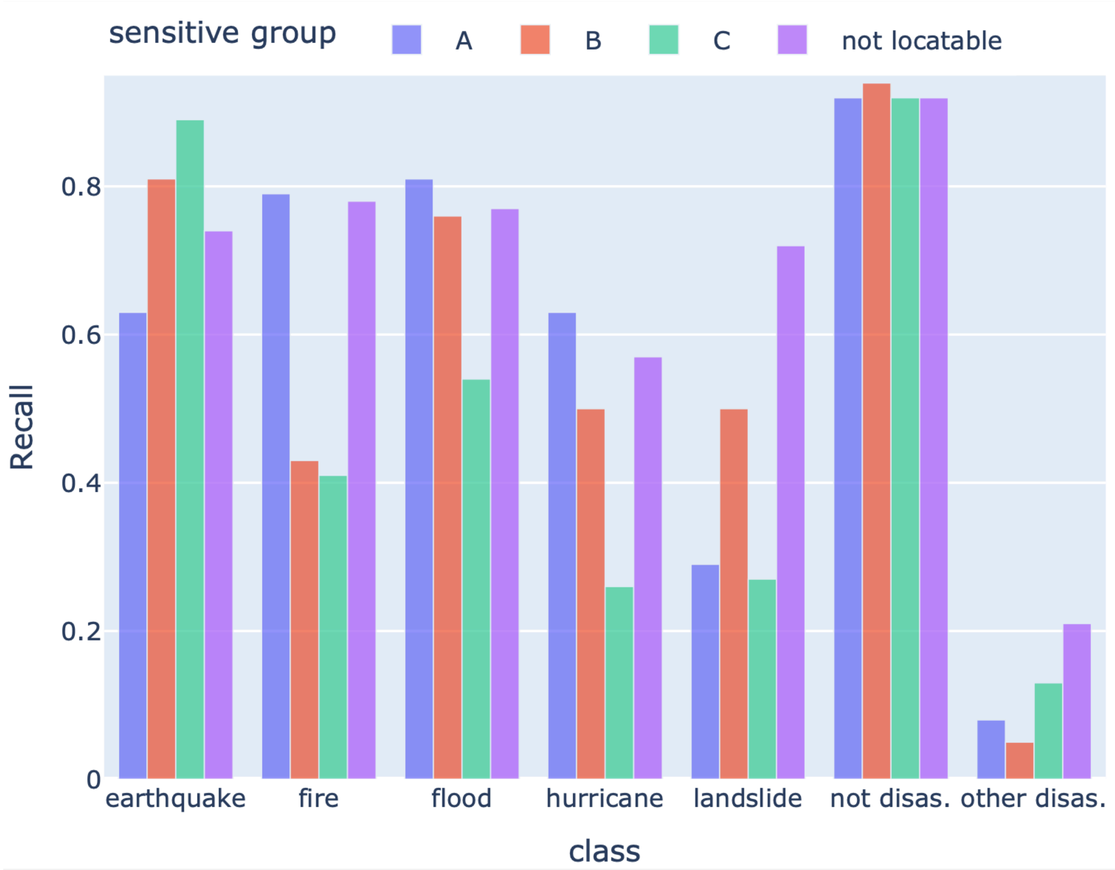
updated schema

field	description
image_id	id of the sample based on the file name
subdataset	name of the dataset where sample comes from
image_path	file path of the sample
disaster_types	classification task
informative	classification task
humanitarian	classification task
damage_severity	classification task
event name	actual name of the disaster event
countries	all involved countries in the disaster event
sensitive group	assigned sensitive group based on the PCA of the development indexes

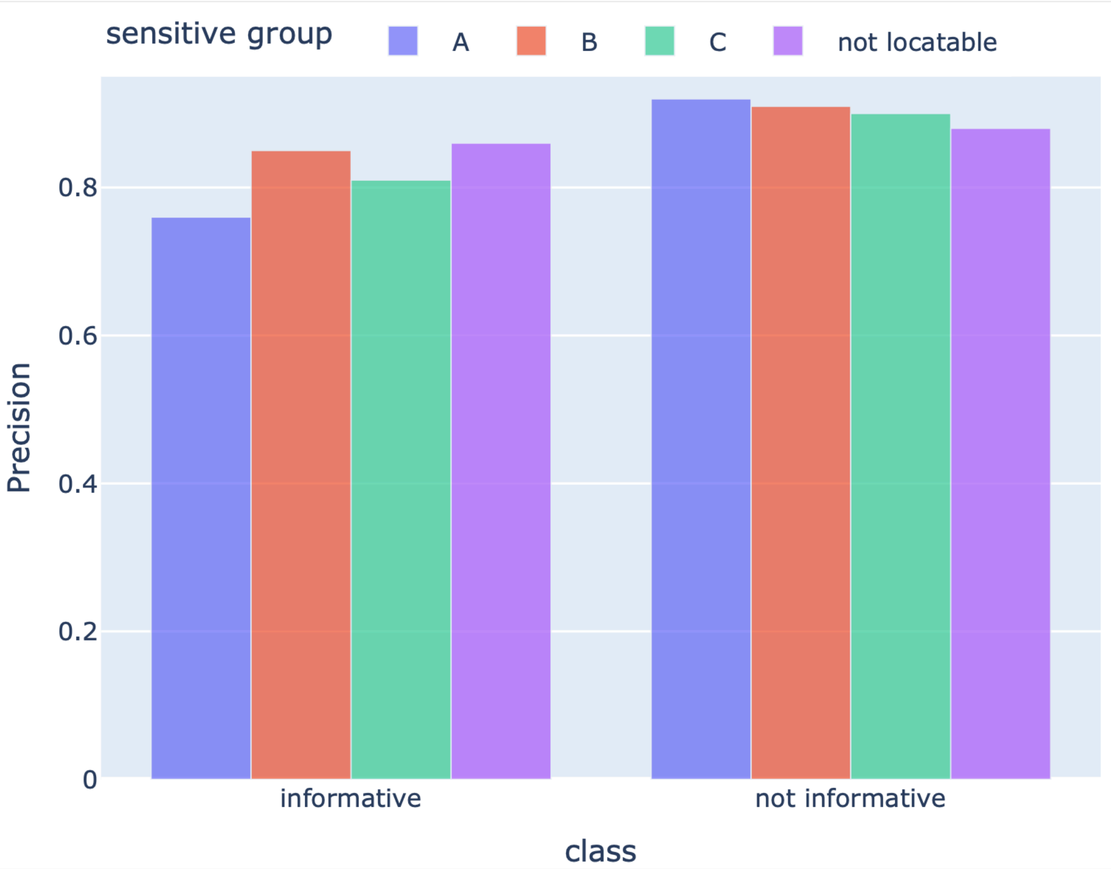
3 results – smaller f1-score parity infringements on task level

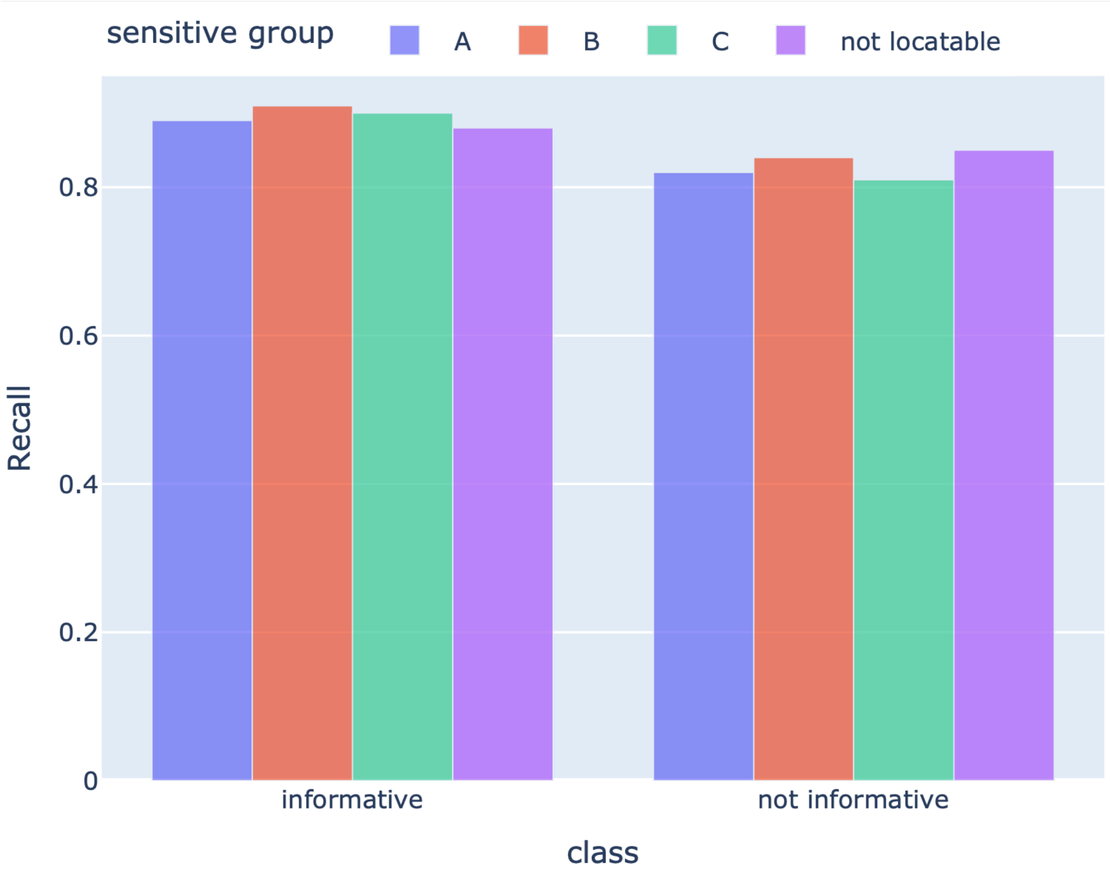


3 results – precision and recall of *disaster types* subtask per sensitive group



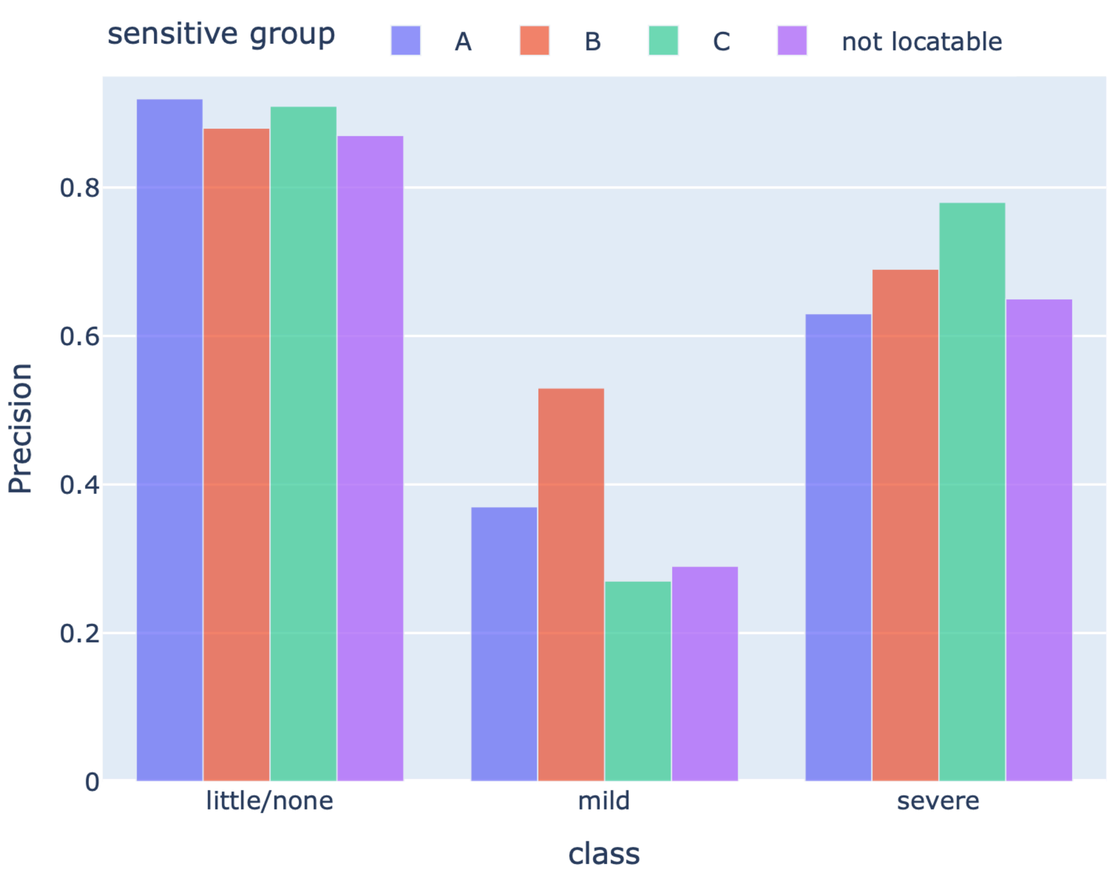
3 results – precision and recall of *informative* subtask per sensitive group

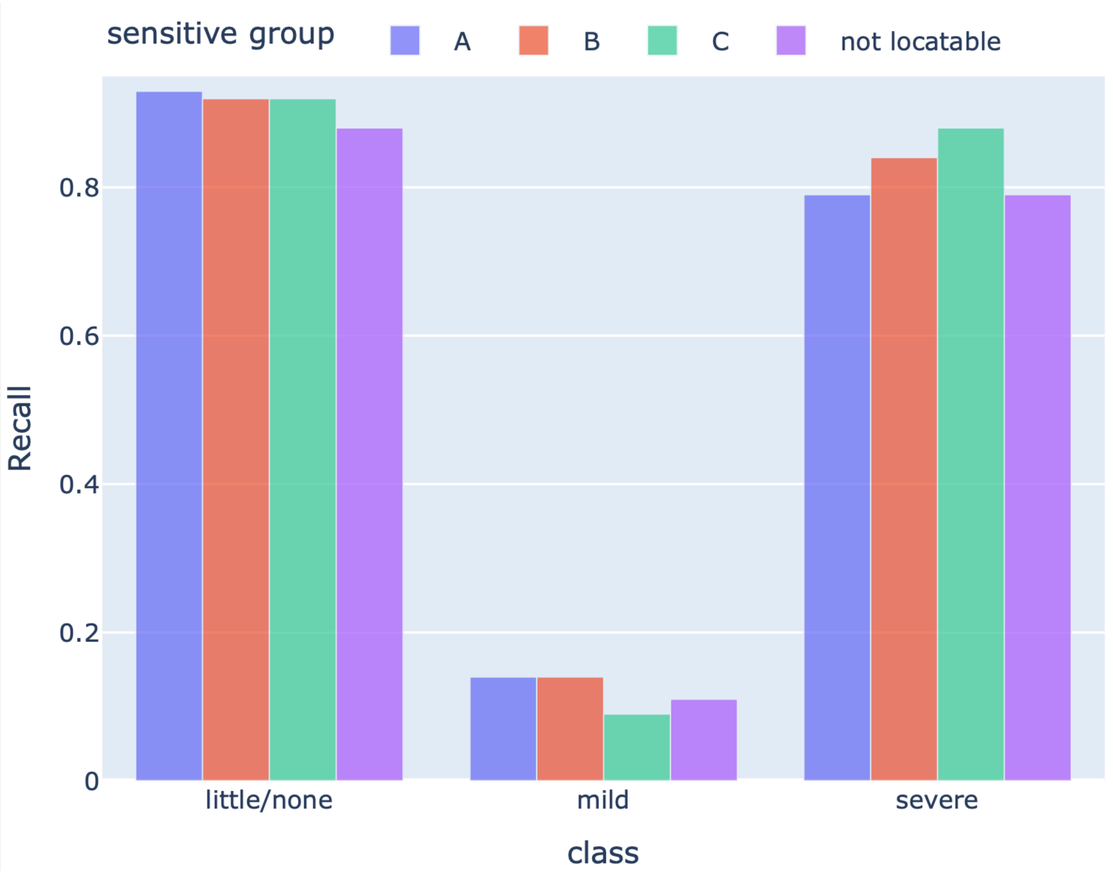




3 results – precision and recall of *humanitarian* subtask per sensitive group

3 results – precision and recall of *damage severity* subtask per sensitive group





3 results - **comparing predictive parity and equal opportunity infringements**

precision deviations from best performing group

task	subtask	A	B	C
disaster types	earthquake			
	fire			
	flood			
	hurricane			
	landslide			
	other disaster			
	not disaster			
informative	yes			
	no			
humanitarian	injured or dead			
	infrastructure			
	rescue or donation			
	not humanitarian			
damage severity	little or none			
	mild			
	severe			

recall deviations from best performing group

A	B	C	subtask	task
			earthquake	disaster types
			fire	
			flood	
			hurricane	
			landslide	
			other disaster	
			not disaster	
			yes	informative
			no	
			injured or dead	humanitarian
			infrastructure	
			rescue or donation	
			not humanitarian	
			little or none	damage severity
			mild	
			severe	

0.41 ... 1 (tremendous)
0.31 .. 0.4 (huge)
0.21 . .0.3 (large)
0.11 .. 0.2 (moderate)
0.04 .. 0.1 (low)
0 .. 0.03 (no)

comparable pattern between predictive parity and equal opportunity

although equal opportunity is roughly about 10% less infringed

strongest disadvantages per group:

A: earthquake, landslide, other disaster, injured or dead

B: fire, other disaster, injured or dead

C: fire, flood, hurricane, rescue or donation, mild damage severity

3 results - **other observations**

low sample sizes are often correlated to fairness infringements, but not for every subtask

e.g.:

earthquake: B and C are equally good with 349 and 929 samples

flood: B has a 15% points better performance with 339 samples over A with 480 samples

3 results - **other observations**

low sample sizes are often correlated to fairness infringements, but not for every subtask

e.g.:
earthquake: B and C are equally good with 349 and 929 samples

flood: B has a 15% points better performance with 339 samples over A with 480 samples

misclassifications often are made towards the negative class of the subclass

3 results - **other observations**

low sample sizes are often correlated to fairness infringements, but not for every subtask

e.g.:
earthquake: B and C are equally good with 349 and 929 samples

flood: B has a 15% points better performance with 339 samples over A with 480 samples

misclassifications often are made towards the negative class of the subclass

some subtasks get especially confused with each other

hurricanes with *earthquakes* and *floods,*

injured/dead with *infrastructure damage,*

damage severity mild with *severe*

3 results - **other observations**

low sample sizes are often correlated to fairness infringements, but not for every subtask

e.g.:
earthquake: B and C are equally good with 349 and 929 samples

flood: B has a 15% points better performance with 339 samples over A with 480 samples

different subtasks need different amount of image to be learned by the classifier

e.g.:
hurricanes and *damage severity mild* and *severe* have relatively high sample amounts but low scores

misclassifications often are made towards the negative class of the subclass

fairness problems are not tied to the difficulty of the tasks

fairness infringements appear both on subtasks that trend to higher scores as well as on subtasks that tend to score bad

some subtasks get especially confused with each other

hurricanes with *earthquakes* and *floods,*

injured/dead with *infrastructure damage,*

damage severity mild with *severe*

predictive performances and fairness infringements of negative samples are distinct to positive classes

negative classes:
- have high samples sizes
- low fairness infringements
- seem easier to learn
- their image quality (visuality) across subtasks seems homogenous

4 evaluation – **interpretation of fairness infringements**

predictive parity
is more infringed than
equal opportunity

disadvantaged group w.r.t. **predictive parity** tends to more **false alarms** than other groups

disadvantaged group w.r.t. **equal opportunity** tends to have more **misses**

4 evaluation – **interpretation of fairness infringements**

predictive parity
is more infringed than
equal opportunity

disadvantaged group w.r.t. **predictive parity** tends to more **false alarms** than other groups

disadvantaged group w.r.t. **equal opportunity** tends to have more **misses**

dissimilar impact of
fairness shortcomings for
positive and negative classes

predictive parity is important for negative classes

equal opportunity is important for positive classes

4 evaluation – **interpretation of fairness infringements**

predictive parity is more infringed than equal opportunity

disadvantaged group w.r.t. **predictive parity** tends to more **false alarms** than other groups

disadvantaged group w.r.t. **equal opportunity** tends to have more **misses**

dissimilar impact of fairness shortcomings for positive and negative classes

predictive parity is important for negative classes

equal opportunity is important for positive classes

no systematic discrimination of one country by development indexes

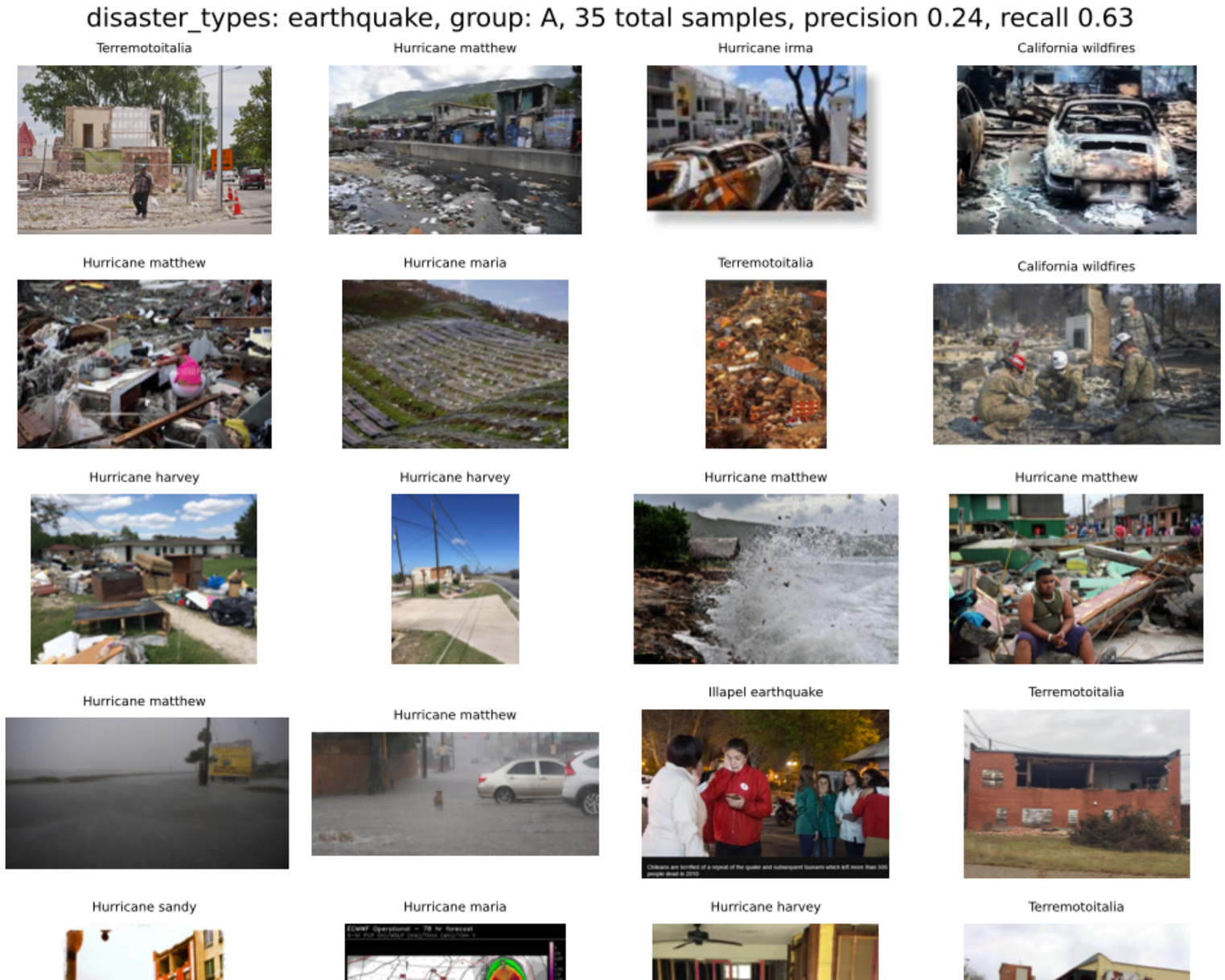
no underprivileged group, no pattern

lower development indexes do not relate to disadvantages within the classifier

biases are based on different subtask and group

groupings by development indexes rather explain the different expressions of disasters regarding being documented in social media by different groups

4 evaluation – labelling is a difficult (and open) task on MEDIC



4 evaluation – **explanation of fairness shortcomings**

predictive parity infringements

main cause:

representation bias (class imbalance)

explanation:

huge amount of negative samples (49% to 71%) contribute mostly to false positives and low precision

especially when subtasks are underrepresented by sample size

4 evaluation – **explanation of fairness shortcomings**

predictive parity infringements

main cause:

representation bias (class imbalance)

explanation:

huge amount of negative samples (49% to 71%) contribute mostly to false positives and low precision

especially when subtasks are underrepresented by sample size

equal opportunity infringements

main cause:

different disaster depictions per group
+
representation bias (class imbalance)

explanation:

is based on misses of own subclass (recall)

class imbalance can amplify difficulties to learn a certain subclass per group

4 evaluation – **explanation of fairness shortcomings**

predictive parity infringements

main cause:

representation bias (class imbalance)

explanation:

huge amount of negative samples (49% to 71%) contribute mostly to false positives and low precision

especially when subtasks are underrepresented by sample size

equal opportunity infringements

main cause:

different disaster depictions per group
+
representation bias (class imbalance)

explanation:

is based on misses of own subclass (recall)

class imbalance can amplify difficulties to learn a certain subclass per group

noisy labels

can amplify representation bias and inflict lower predictive performances

4 evaluation – **limitations of the method**

unambiguous localisation of the samples is difficult for **disaster events that span multiple countries**

4 evaluation – **limitations of the method**

unambiguous localisation of the samples is difficult for **disaster events that span multiple countries**

oversea and unincorporated regions might not be represented well by development indexes

4 evaluation – **limitations of the method**

unambiguous localisation of the samples is difficult for **disaster events that span multiple countries**

oversea and unincorporated regions might not be represented well by development indexes

localisation of samples in MEDIC is based on **heuristics**

4 evaluation – **limitations of the method**

unambiguous localisation of the samples is difficult for **disaster events that span multiple countries**

oversea and unincorporated regions might not be represented well by development indexes

localisation of samples in MEDIC is based on **heuristics**

the nature of development indexes is to **subsume fine-grained information** into single representation

5 outlook

exploring diversity of disaster event depictions

model interpretation with XAI, LIME, Grad-Cam, etc.

examining **relationship of images to sensitive groups or development indexes** (e.g. by regression with image clusters/prototypes)

5 outlook

exploring diversity of disaster event depictions

model interpretation with XAI, LIME, Grad-Cam, etc.

examining **relationship of images to sensitive groups or development indexes** (e.g. by regression with image clusters/prototypes)

using more transnational data that is related to disasters

Gender Inequality Index, Gini-Index, Multi-dimensional Poverty Index, ...

INFORM data
(over 100 indices related to DRM on country level)

5 outlook

exploring diversity of disaster event depictions

model interpretation with XAI, LIME, Grad-Cam, etc.

examining **relationship of images to sensitive groups or development indexes** (e.g. by regression with image clusters/prototypes)

using more transnational data that is related to disasters

Gender Inequality Index, Gini-Index, Multi-dimensional Poverty Index, ...

INFORM data
(over 100 indices related to DRM on country level)

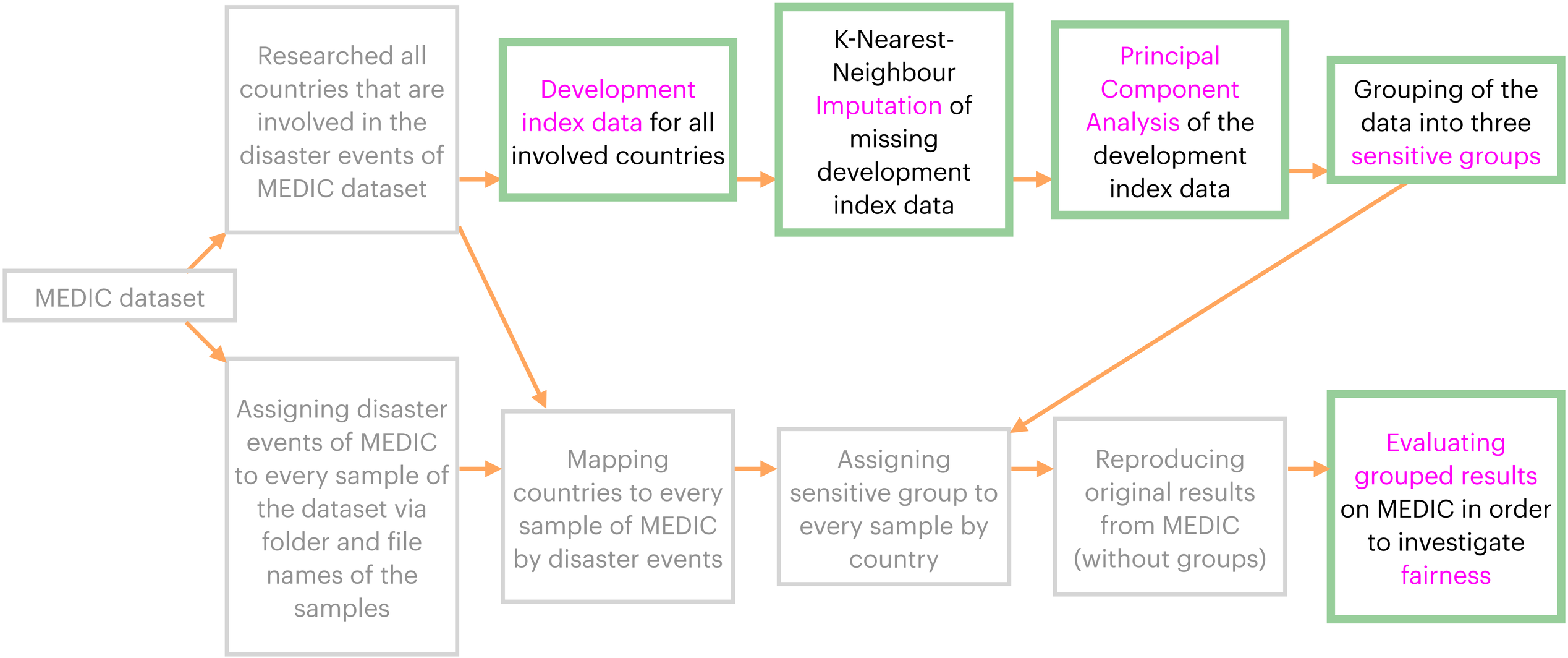
test robustness of method

ablation tests against:

- development index data
- methods (imputation, clustering dimensionality reduction, grouping)

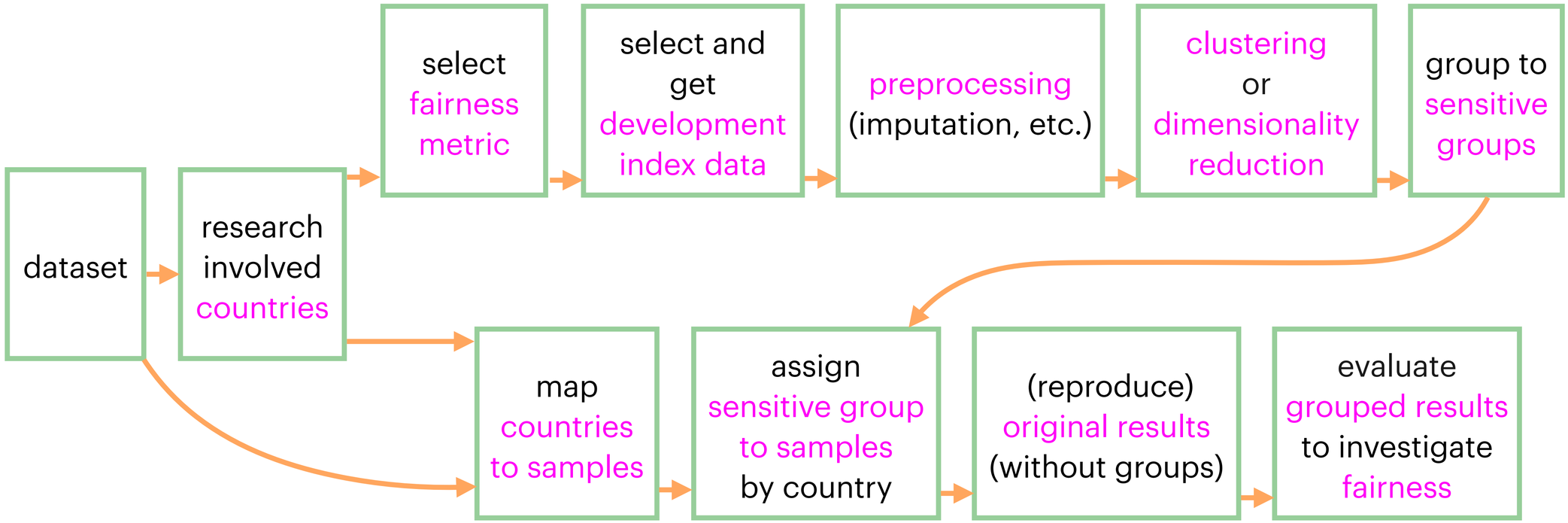
test method on other datasets
(*ImageNet, Dollar Street, ...*)

5 outlook - interchangeable parts of the method



5 outlook –

framework for investigating group fairness on transnational level with development indexes



Thank you!

references & keywords

Alam et al., 2021, MEDIC: A Multi-Task Learning Dataset for Disaster Image Classification

Imran et al., 2014 AIDR: Artificial Intelligence for Disaster Response

Li et al., 2013 Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr

Zou et al., 2019 Social and Geographical Disparities in Twitter Use during Hurricane Harvey

Wang et al., 2019 Are Vulnerable Communities Digitally Left behind in Social Responses to Natural Disasters? An Evidence from Hurricane Sandy with Twitter Data

Dargin et al., 2021 Vulnerable Populations and Social Media Use in Disasters: Uncovering the Digital Divide in Three Major U.S. Hurricanes

Zhu et al., 2021 Temporal, Spatial, and Socioeconomic Dynamics in Social Media Thematic Emphases during Typhoon Mangkhut

Shankar et al., 2017 No Classification without Repre-sentation: Assessing Geodiversity Issues in Open Data Sets for the Developing World

DeVries et al., 2019 Does Object Recognition Work for Everyone?

Goya et al., 2022 Fairness Indicators for Systematic Assessments of Visual Feature Extractors

ML	machine learning
DRM	disaster risk management
PCA	Principal Component Analysis
PCs	principal components

appendix – precision and support table of subtasks

task	subcategory	correlation	precision				# samples			
			A	B	C	n.l.	A	B	C	n.l.
disaster type	earthquake	0.74	0.24	0.8	0.82	0.46	94	349	929	423
	fire	0.90	0.84	0.5	0.38	0.79	266	7	29	388
	flood	0.79	0.76	0.91	0.48	0.8	480	339	41	455
	hurricane	0.82	0.69	0.52	0.29	0.53	1066	105	53	294
	landslide	0.98	0.15	0.25	0.12	0.78	14	4	11	302
	not disaster	0.28	0.9	0.87	0.88	0.83	3569	1004	1849	2463
	other disaster	0.90	0.31	0.12	0.58	0.77	191	37	238	688
informative	yes	-0.19	0.76	0.85	0.81	0.86	2234	914	1505	2553
	no	0.14	0.92	0.91	0.9	0.88	3446	931	1645	2460
humanitarian	injured or dead people	0.99	0.23	0.29	0.65	0.24	94	89	3 61	95
	infrastructure damage	-0.41	0.78	0.77	0.82	0.76	1633	618	976	1997
	not humanitarian	0.89	0.91	0.87	0.87	0.88	3658	967	1675	2845
	rescue or donation	0.60	0.49	0.65	0.26	0.21	295	171	138	76
damage severity	little or none	0.43	0.92	0.88	0.91	0.87	4080	1115	2008	3049
	mild	-0.33	0.37	0.53	0.27	0.29	542	226	219	540
	severe	-0.36	0.63	0.69	0.78	0.65	1058	504	923	1424

appendix – recall and support table of subtasks

task	subcategory	correlation	recall				# samples			
			A	B	C	n.l.	A	B	C	n.l.
disaster type	earthquake	0.90	0.63	0.81	0.89	0.74	94	349	929	423
	fire	0.95	0.79	0.43	0.41	0.78	266	7	29	388
	flood	0.98	0.81	0.76	0.54	0.77	480	339	41	455
	hurricane	0.71	0.63	0.5	0.26	0.57	1066	105	53	294
	landslide	0.86	0.29	0.5	0.27	0.72	14	4	11	302
	not disaster	-0.75	0.92	0.94	0.92	0.92	3569	1004	1849	2463
	other disaster	0.97	0.08	0.05	0.13	0.21	191	37	238	688
informative	yes	-0.99	0.89	0.91	0.9	0.88	2234	914	1505	2553
	no	-0.15	0.82	0.84	0.81	0.85	3446	931	1645	2460
humanitarian	injured or dead people	0.96	0.17	0.22	0.54	0.09	94	89	3 61	95
	infrastructure damage	-0.25	0.86	0.84	0.9	0.84	1633	618	976	1997
	not humanitarian	-0.31	0.9	0.91	0.88	0.84	3658	967	1675	2845
	rescue or donation	0.76	0.31	0.34	0.17	0.11	295	171	138	76
damage severity	little or none	-0.05	0.93	0.92	0.92	0.88	4080	1115	2008	3049
	mild	0.25	0.14	0.14	0.09	0.11	542	226	219	540
	severe	-0.57	0.79	0.84	0.88	0.79	1058	504	923	1424

appendix – f1-score and support table of subtasks

task	subcategory	correlation	F1-Score A	F1-Score B	F1-Score C	F1-Score n.l.	# samples A	# samples B	# samples C	# samples n.l.
disaster type	earthquake	0.79	0.35	0.8	0.86	0.57	94	349	929	423
	fire	0.93	0.81	0.46	0.39	0.78	266	7	29	388
	flood	0.89	0.78	0.83	0.51	0.78	480	339	41	455
	hurricane	0.77	0.66	0.51	0.27	0.55	1066	105	53	294
	landslide	0.96	0.2	0.33	0.17	0.74	14	4	11	302
	not disaster	0.11	0.91	0.9	0.9	0.87	3569	1004	1849	2463
	other disaster	0.96	0.13	0.07	0.21	0.33	191	37	238	688
informative	yes	-0.47	0.82	0.88	0.86	0.87	2234	914	1505	2553
	no	0.20	0.87	0.87	0.85	0.86	3446	931	1645	2460
humanitarian	injured or dead people	0.97	0.2	0.25	0.59	0.14	94	89	3 61	95
	infrastructure damage	-0.41	0.82	0.81	0.86	0.8	1633	618	976	1997
	not humanitarian	0.29	0.91	0.89	0.87	0.86	3658	967	1675	2845
	rescue or donation	0.70	0.38	0.45	0.21	0.14	295	171	138	76
damage severity	little or none	0.19	0.93	0.9	0.92	0.87	4080	1115	2008	3049
	mild	0.09	0.2	0.22	0.13	0.16	542	226	219	540
	severe	-0.46	0.7	0.76	0.83	0.71	1058	504	923	1424

appendix –

confusion matrices: disaster types

support

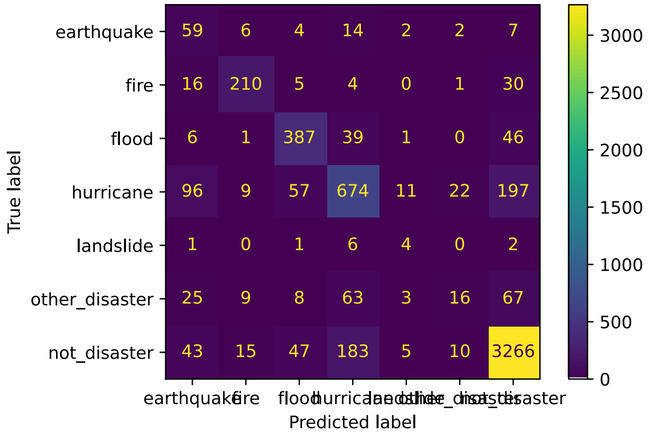
precision

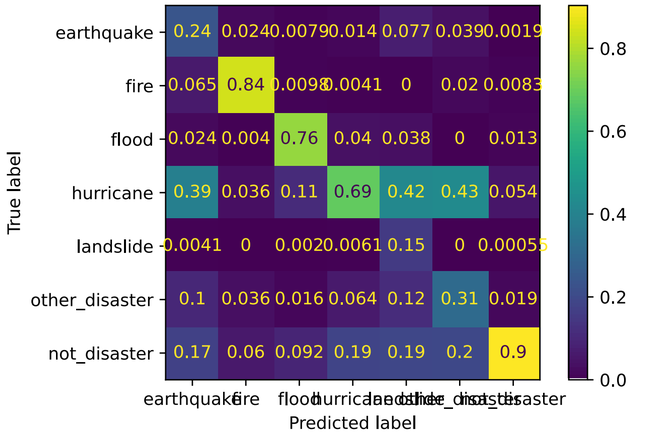
(read columns / top to bottom)

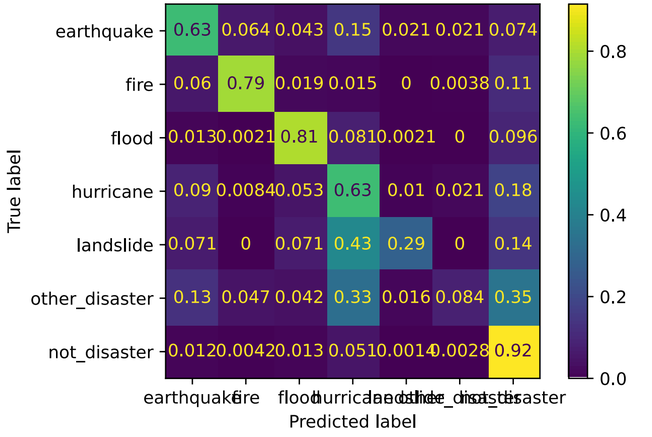
recall

(read rows / left to right)

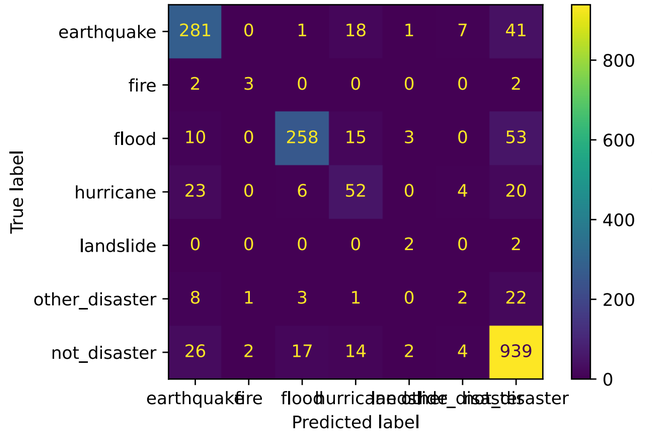
A

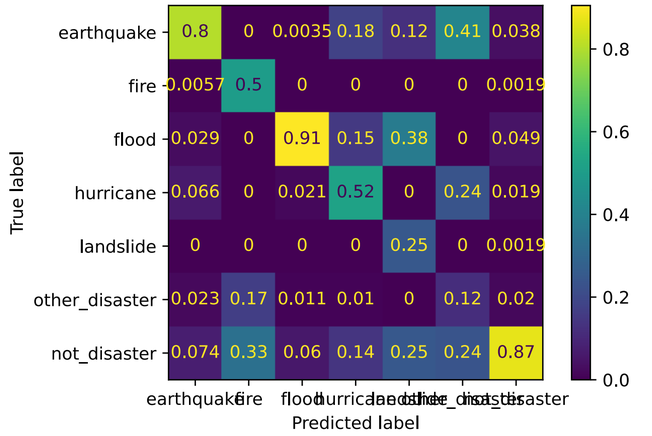


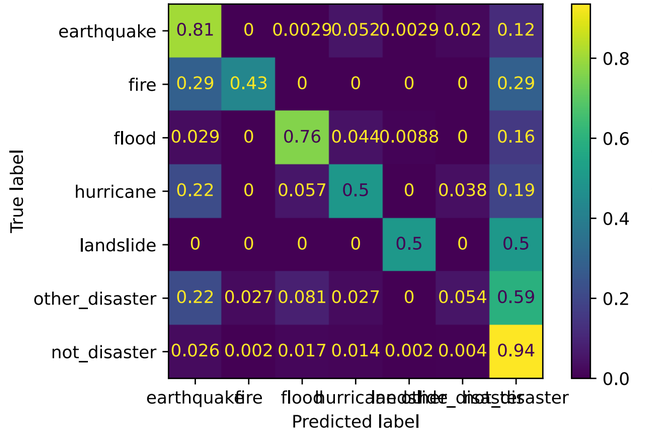




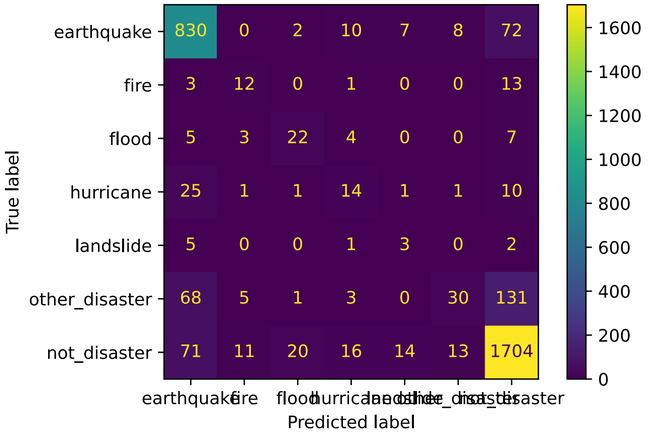
B

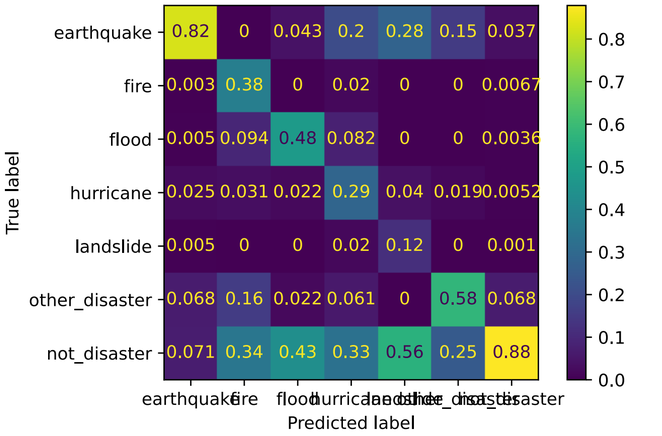


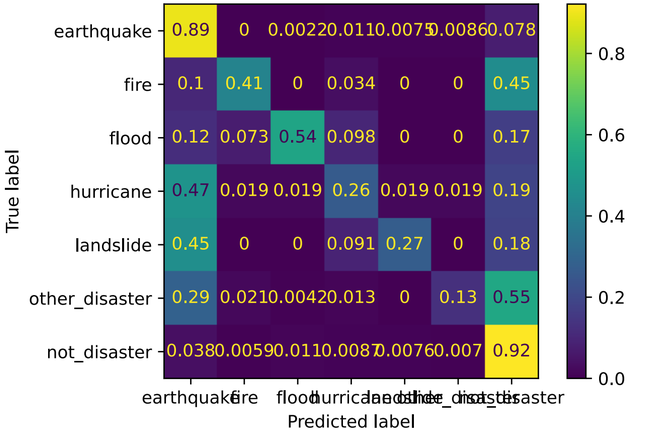




C



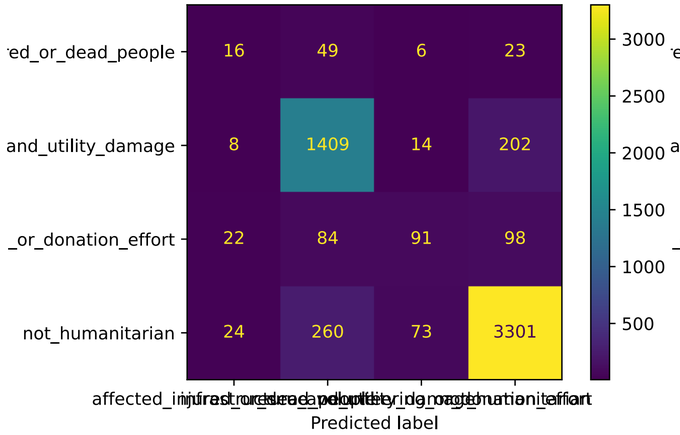




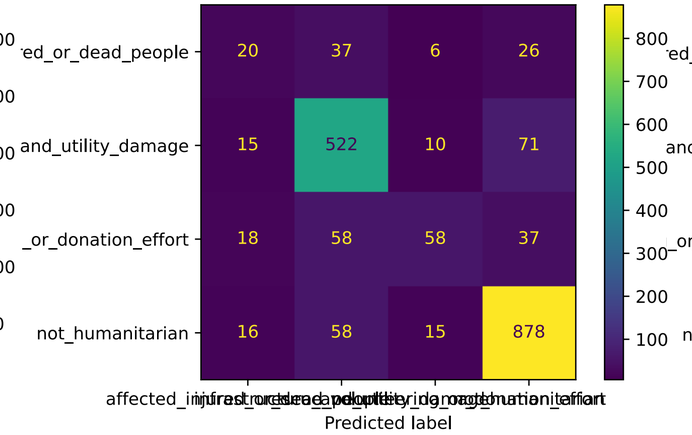
appendix –

confusion matrices: humanitarian

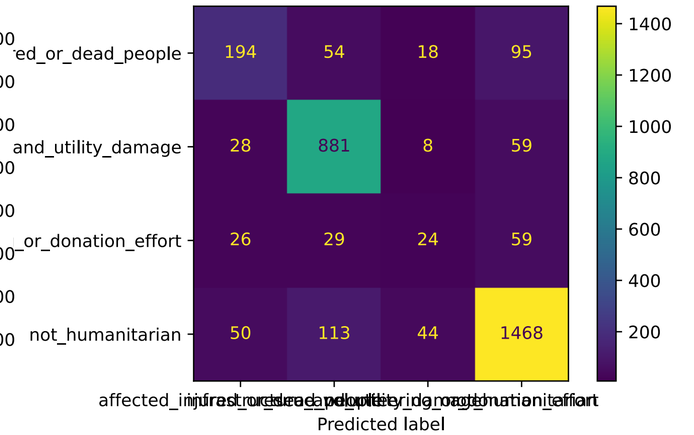
A



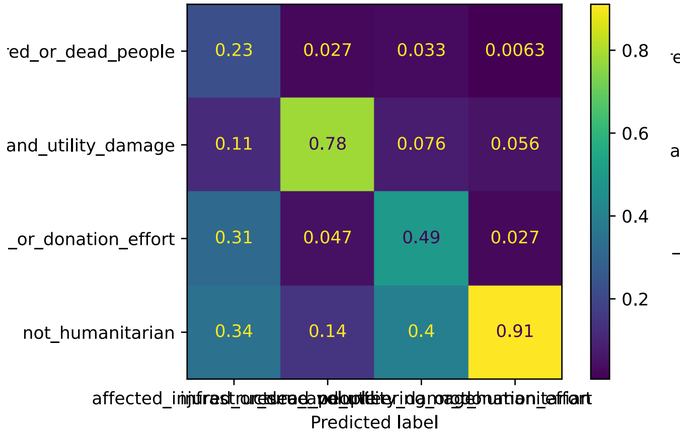
B

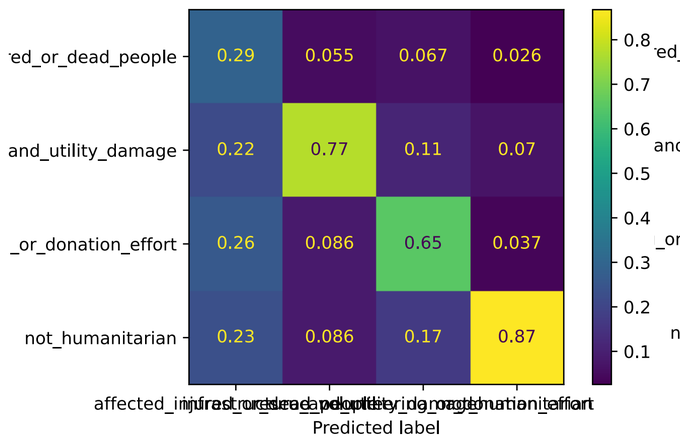


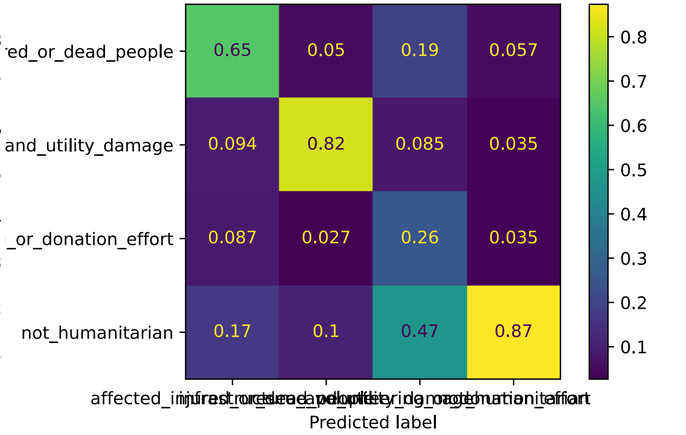
C



support

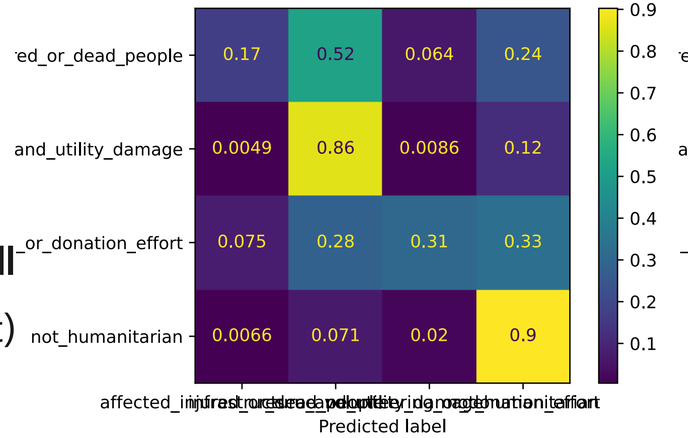


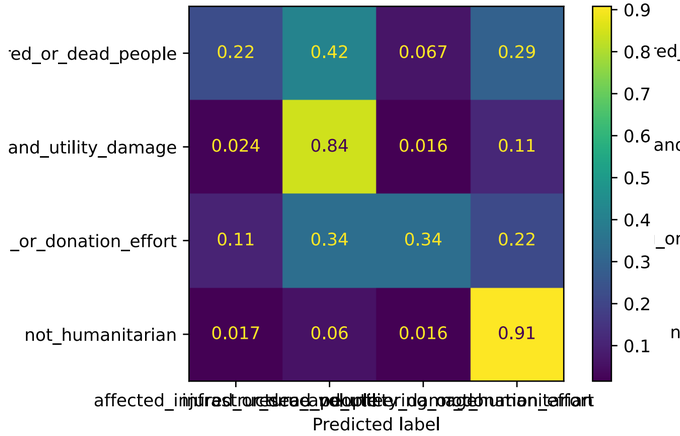


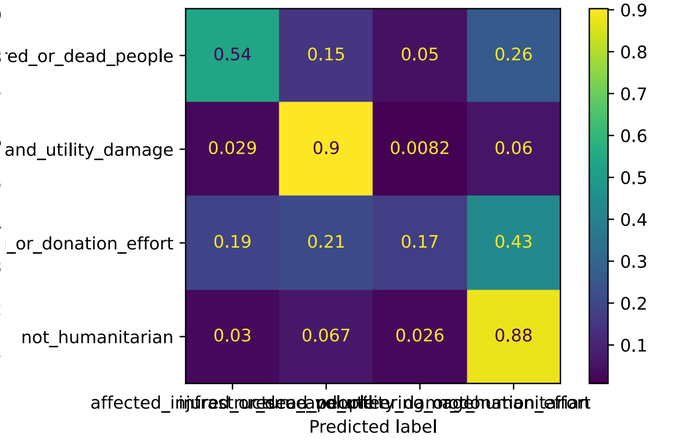


precision

(read columns / top to bottom)







recall

(read rows / left to right)

appendix –

confusion matrices: damage severity

support

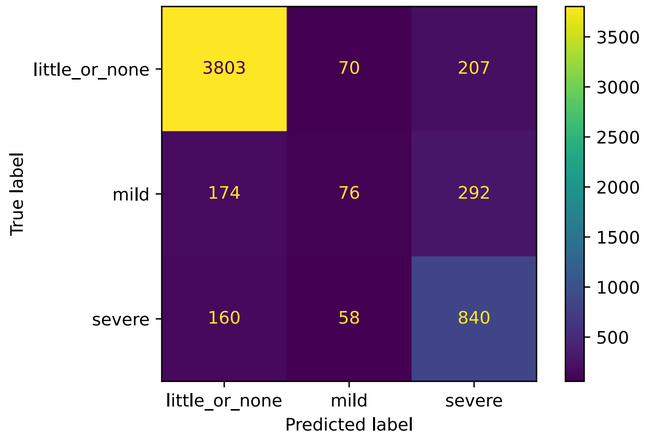
precision

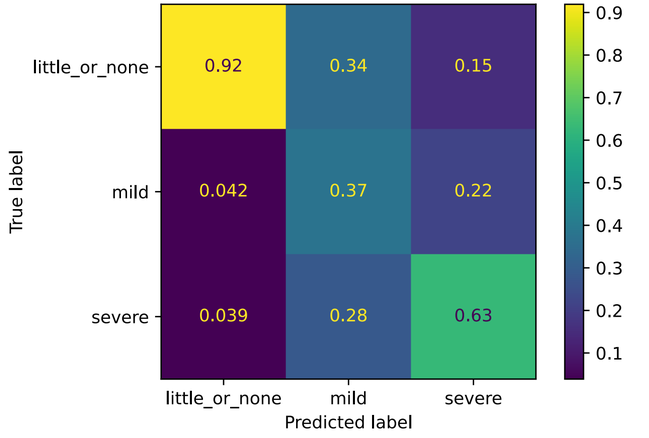
(read columns / top to bottom)

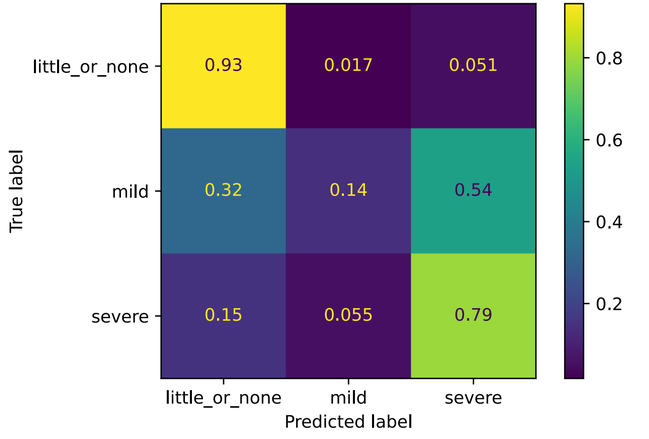
recall

(read rows / left to right)

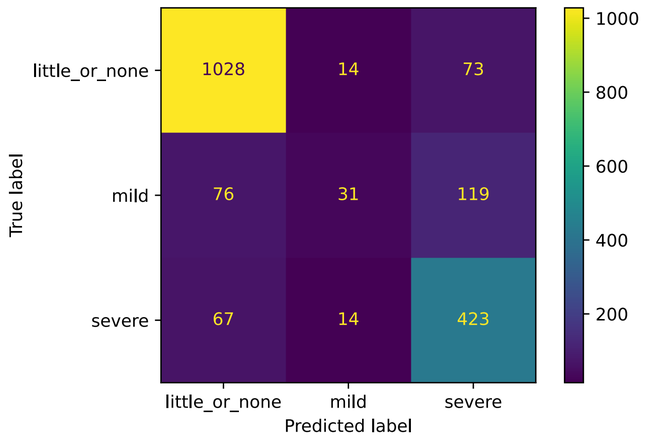
A

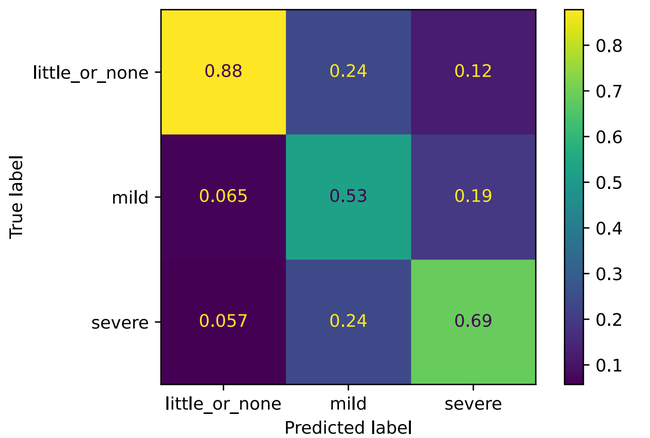


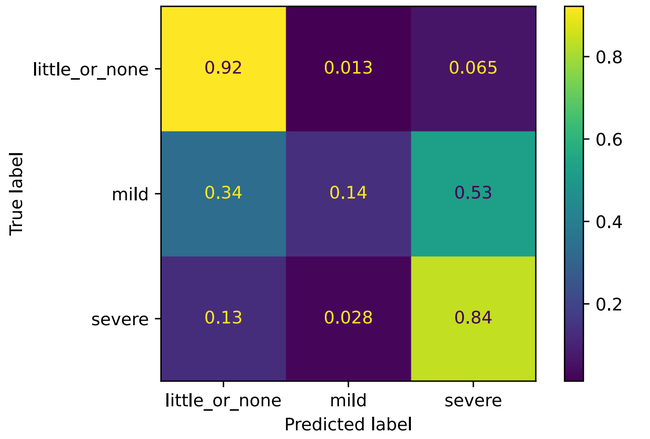




B







C