

Emerging Data Landscapes in M&E

Getting Practical with Text Mining for Evaluation:

How to Make Sense of Project Information

15 April 2021



**EUROPEAN
EVALUATION
SOCIETY**

- Welcome! Feel free to introduce yourself on the chat
- We invite you to ask questions during the presentations in the chat box and Q&A box, so the moderator and presenters can respond
- In the Q&A, you can “up vote” certain questions by clicking the “like” (thumbs up) next to a question

Welcome

Daniele Lamarque, EES President

Dr. Sven Harten, EES TWG 6

Session	Speakers / Facilitators
Introduction	Daniele Lamarque, EES President Dr. Sven Harten, DEval & EES TWG 6
Session 1: GEOTEX: A tool to geocode project locations based on project documents	Thomas Wencker, Evaluator, DEval
Session 2: Automated project classification using textual information in aid statistics: an example from the Women, Peace and Security portfolio	Javier Fabra-Mata, PhD, Senior Advisor, Evaluation Department, Norad
Session 3: Using Machine Learning for Analysing Human Rights Mainstreaming in Development Project Documents	Lea Smidt, Evaluator, DEval
Q&A	Facilitated by Estelle Raimondo, EES Board Member
Wrap up	Alena Lappo, EES Board Member

Our question is:

*What is your level of experience in
text mining?*

**Session 1: GEOTEX: A tool to
geocode project locations based on
project documents**

Thomas Wencker, Evaluator, DEval

GeoTex

A tool to geocode project locations based on project documents

Thomas Wencker

German Institute for Development Evaluation

15.04.2021

Why spatial information can improve evaluation

Effects are local and ...

- Phenomena of interest to evaluation vary spatially
 - SDG 10 - socio-economic inequality (Smith & Permanyer 2019)
 - SDG 13 - exposure to natural disasters (Leppert et al. 2018)
 - SDG 16 - violent conflict (Cederman & Gleditsch 2009) & state fragility (Wencker & Verspohl 2019)
- Effects of interventions are local (Dreher & Lohmann 2015; Civellia et al. 2018)

... spatially disaggregated data is available.

- Availability and ease of use of spatially disaggregated data has greatly improved
 - geo-coded survey data (e.g. IPUMS)
 - geo-coded event data (e.g. UCDP-GED, ACLED, HIIK)
 - satellite data (e.g. nightlight)

Yet, a level-of-analysis-gap remains

- Large-N comparisons predominantly rely on the country-year as level of analysis

What can be done to overcome the level-of-analysis gap?

Data is not the problem, ...

- Processes of formulating and implementing development cooperation policies create text
- These texts often include geographical information

... but time and resources are.

- The abundance of text renders close reading very resource heavy

GeoTex as an alternative

- The computer performs (parts of) the tasks of a human coder and automatically detects project locations
- GeoTex is a script-based tool to extract place names from text including geocoding
- Based on existing open-source software (R) and state-of-the-art Natural Language Processing (spaCy)
- Developed in collaboration with Leipzig University

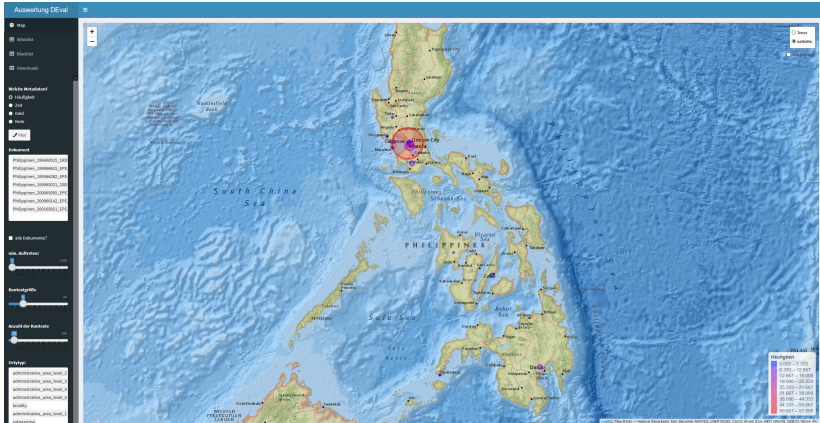
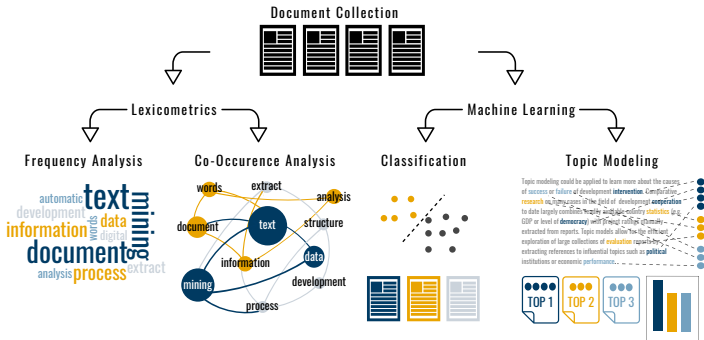


Figure 1: GeoTex Frontend

Named Entity Recognition



Named-entity recognition

- Classification maps document properties to classes
- Classifiers assign classes to unknown text (i.e. automatic labelling of text)
- One application is to identify locations mentioned in texts

Steps

Import and Preprocess

1. Import texts
2. Language detection

Extract locations

3. Extract locations from text using spaCy
4. GeoCoding

Delete false positives

5. Filtering by Type/Class (e.g. shops, tourist sites)
6. Clustering for confidence
7. Filtering against blacklist

Prepare output

8. Enrich locations with external meta information
9. Create, condense and save whitelist
10. Extraction of metadata from texts

Get started

Install Software

- R environment & required packages
- RStudio as front-end
- Python (e.g. Miniconda) & spaCy

Prepare documents

- Check if documents are machine-readable (OCR)
- Move documents to import folder

Download Script

- Run GeoTex script

Validate results

- Draw a sample from documents
- Manually code locations for gold standard
- Check GeoTex results against gold standard

Colombia

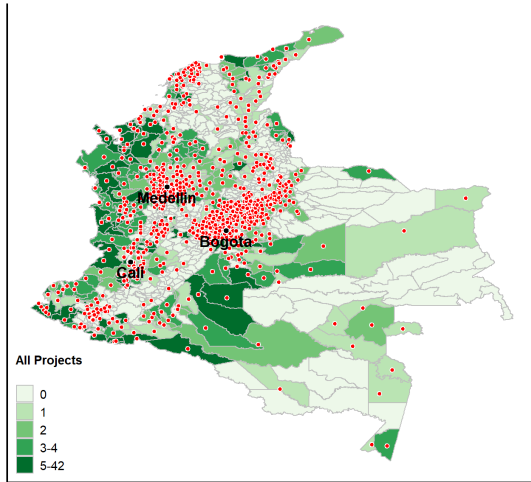


Figure 3: Project locations (Nawrotzki et al. (forthcoming))

Colombia II

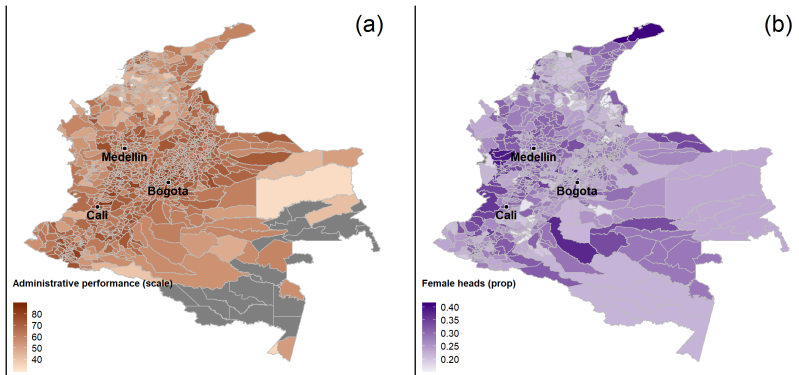


Figure 4: Administrative Performance (DNP-DDDR, 2017) and proportion of female-headed households (IPUMS)

Colombia III

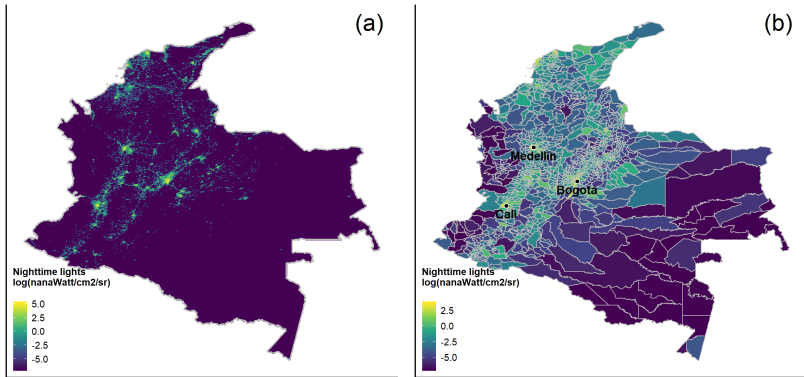


Figure 5: Raster data of nighttime lights from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) made publicly available by the US National Oceanic and Atmospheric Administration (NOAA) (NOAA, 2020)

Citations

Civellia, Horowitz, and Teixeira (2018): Foreign aid and growth: A Sp P-VAR analysis using satellite sub-national data for Uganda. <https://doi.org/10.1016/j.jdeveco.2018.05.001>

Briggs (2017): Does Foreign Aid Target the Poorest? <https://doi.org/10.1017/S0020818316000345>

Dreher and Lohmann 2015: Aid and Growth at the Regional Level. <https://doi.org/10.1093/oxrep/grv026>

Kotsadam, Østby, Rustad, Tollefsen, and Urdal (2018): Development aid and infant mortality. Micro-level evidence from Nigeria. <https://doi.org/10.1016/j.worlddev.2017.12.022>

Leppert, Hohfeld, Lech, and Wencker (2018): Impact, Diffusion and Scaling-Up of a Comprehensive Land-Use Planning Approach in the Philippines. From Development Cooperation to National Policies.

Minnesota Population Center (2020) Integrated Public Use Microdata Series, International: Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2020. <https://doi.org/10.18128/D020.V7.3>

Nawrotzki, Gantner, Balzer, Wencker, and Brüntrup-Seidemann (forthcoming): Strategic allocation of gender-focused projects in post-conflict regions: Evidence from Colombia.

Niekler and Wencker: Text Mining in Evaluation (2019): Text Mining in Evaluation. <https://bit.ly/3gaANBI>

Raleigh, Linke, Hegre, and Karlsen (2010): Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. <https://doi.org/10.1177/0022343310378914>

Sundberg and Melander (2013) Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research* 50(4).

Smith & Permanyer (2019): The Subnational Human Development Database. <https://doi.org/10.1038/sdata.2019.38>

Tollefsen, Strand, and Buhaug (2012). PRIO-GRID: A unified spatial data structure. <https://doi.org/10.1177/0022343311431287>

Trinn and Wencker (2016), The Heidelberg Approach to Conflict Research. <https://doi.org/10.1057/s41304-016-0093-4>

Wencker, Kahmann, Niekler, Nawrotzki, and Islam (forthcoming): GeoTex. A tool to geocode project locations based on project documents

Wencker and Verspohl (2019): German Development Cooperation in Fragile Contexts. <https://bit.ly/2OSfkIW>

**Session 2: Automated project
classification using textual information
in aid statistics: an example from the
Women, Peace and Security portfolio**

Javier Fabra-Mata, PhD, Senior Advisor,
Evaluation Department Norad

EVALUATION DEPARTMENT

Automated project classification using textual information in aid statistics - Norway's Women, Peace and Security portfolio

Javier Fabra-Mata, PhD



Mi Nge Lay/Mon Cetana Organisation



A bit of context

- **The evaluation** - An ongoing evaluation of Norway's support to the Women, Peace and Security (WPS) agenda, 2000 – 2020.
Its scope is limited to the participation of women in peace efforts.
- **Statistics** - OECD code lists are used by donor countries to report on their development aid flows. These codes are used to classify information in the DAC databases.
- **The problem** - needles in a haystack.
A large universe of potential interventions. There is no OECD statistical code for “women’s participation in peace”, only proxies. *How do we find the right projects in the portfolio?*

The problem

Using OECD DAC coding system

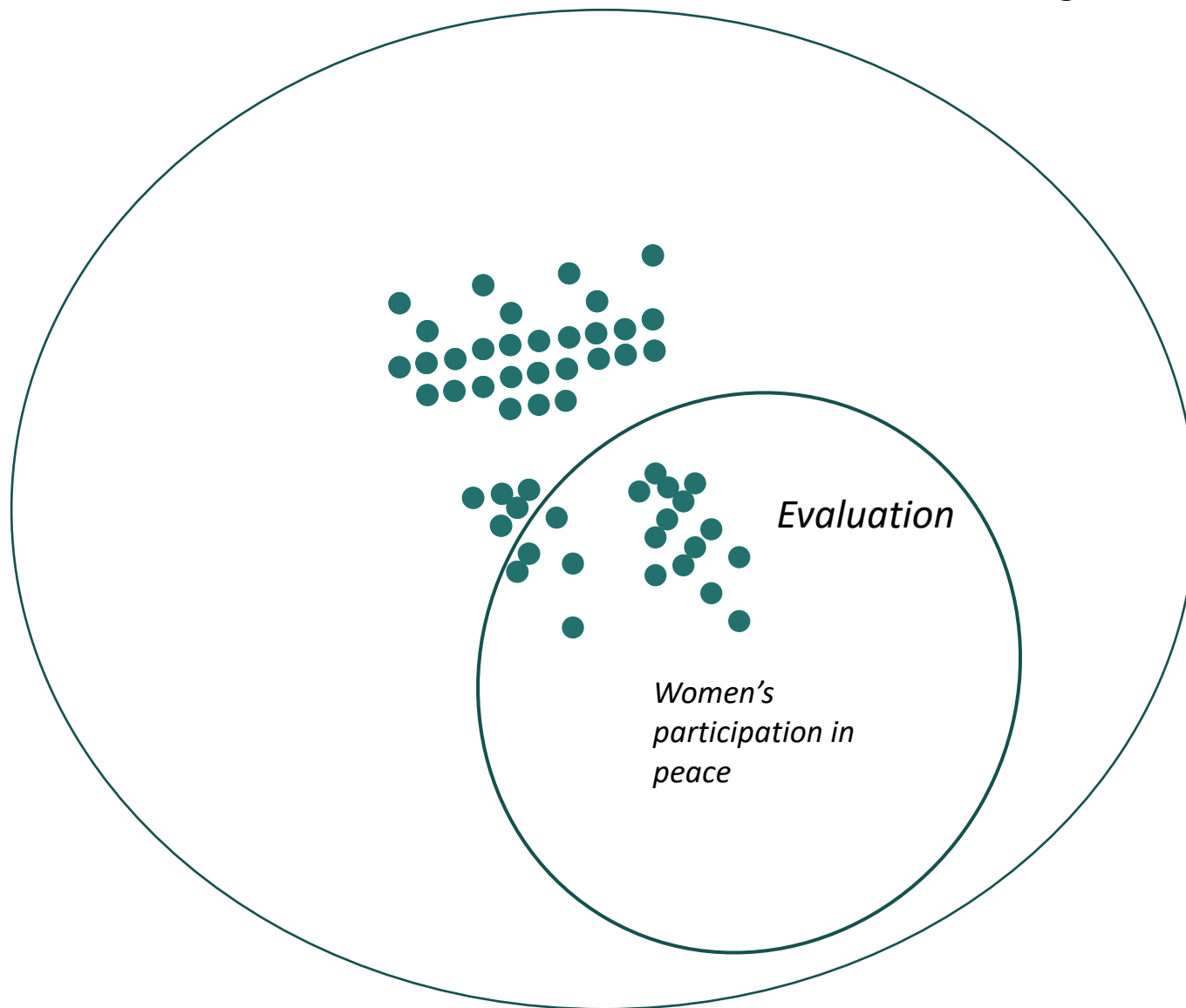
*Women,
peace
and
security*



The problem

Using **OECD DAC coding system**

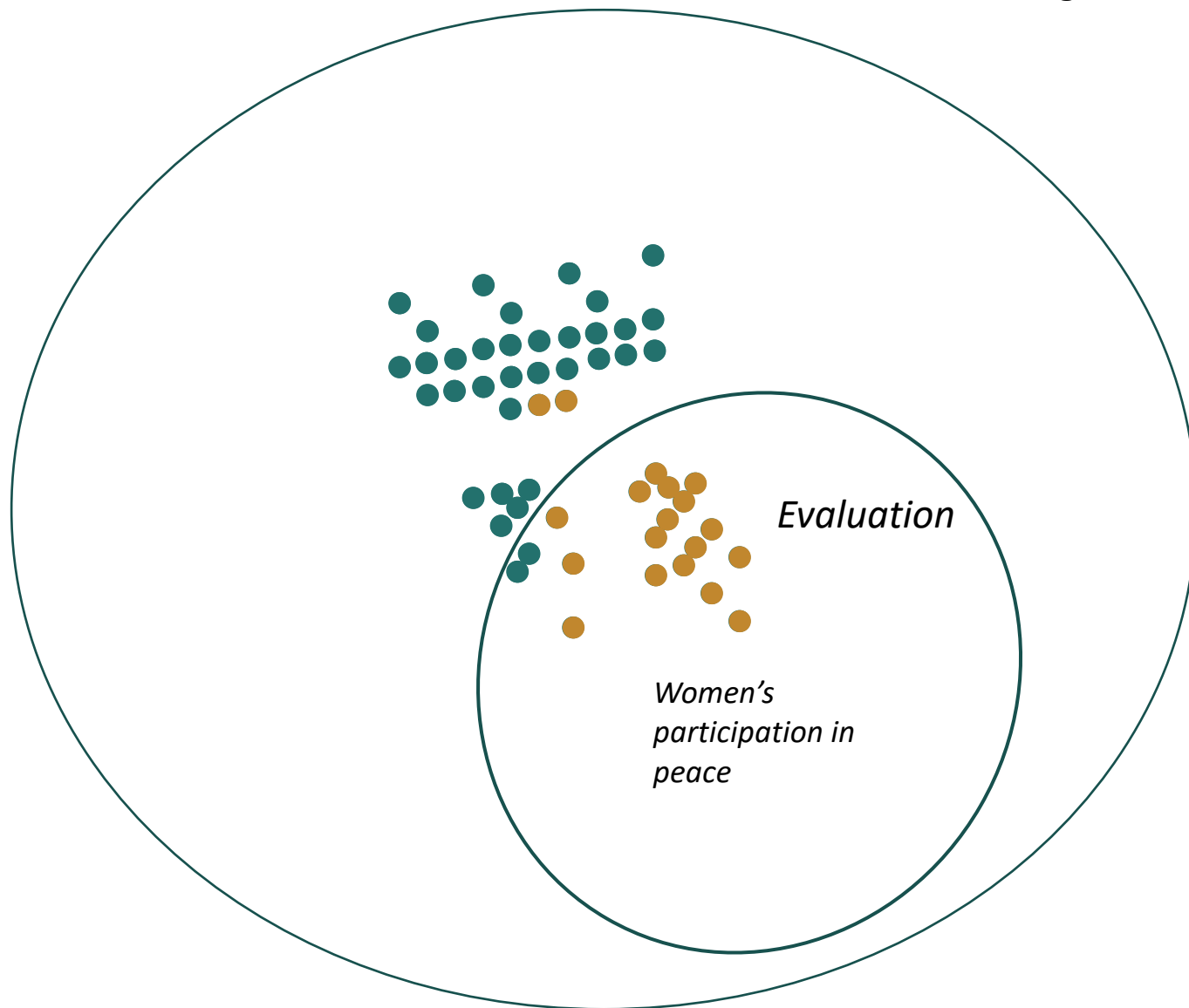
*Women,
peace
and
security*



The problem

Using OECD DAC coding system

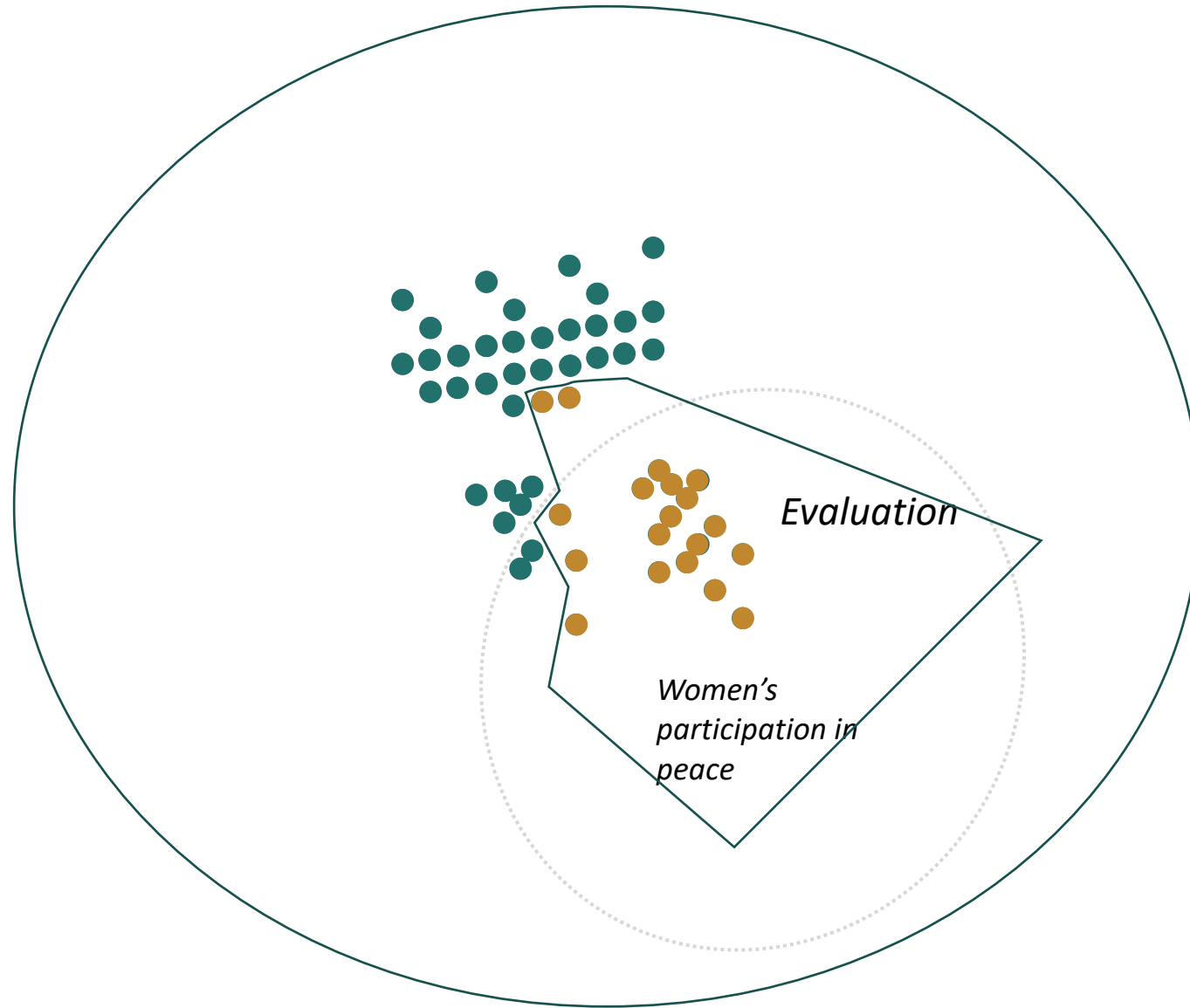
*Women,
peace
and
security*



A solution

Using **textual descriptions**

*Women,
peace
and
security*

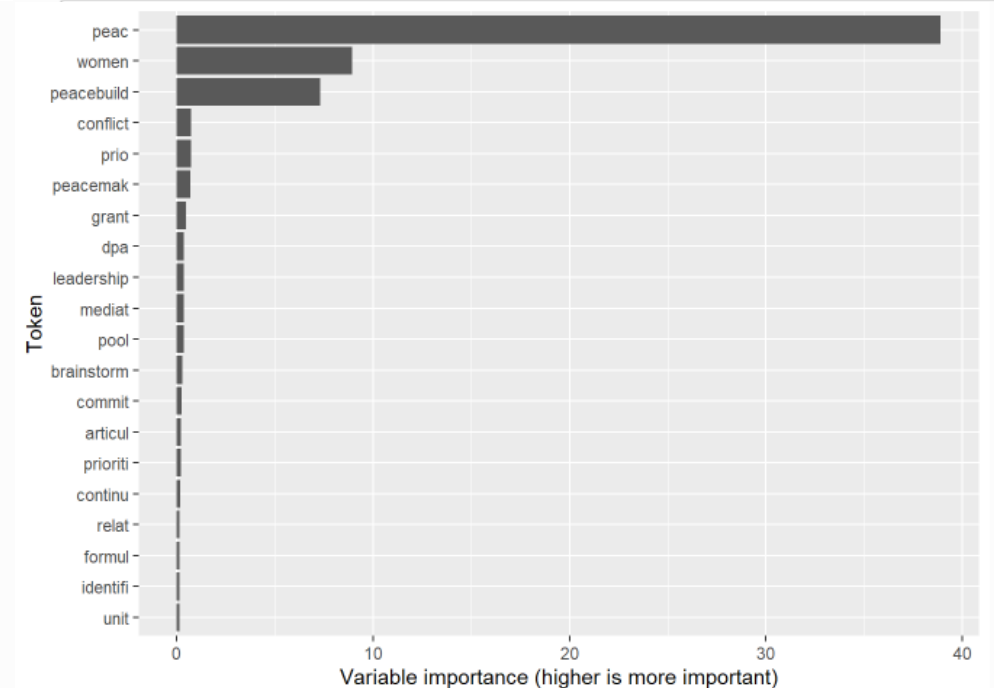
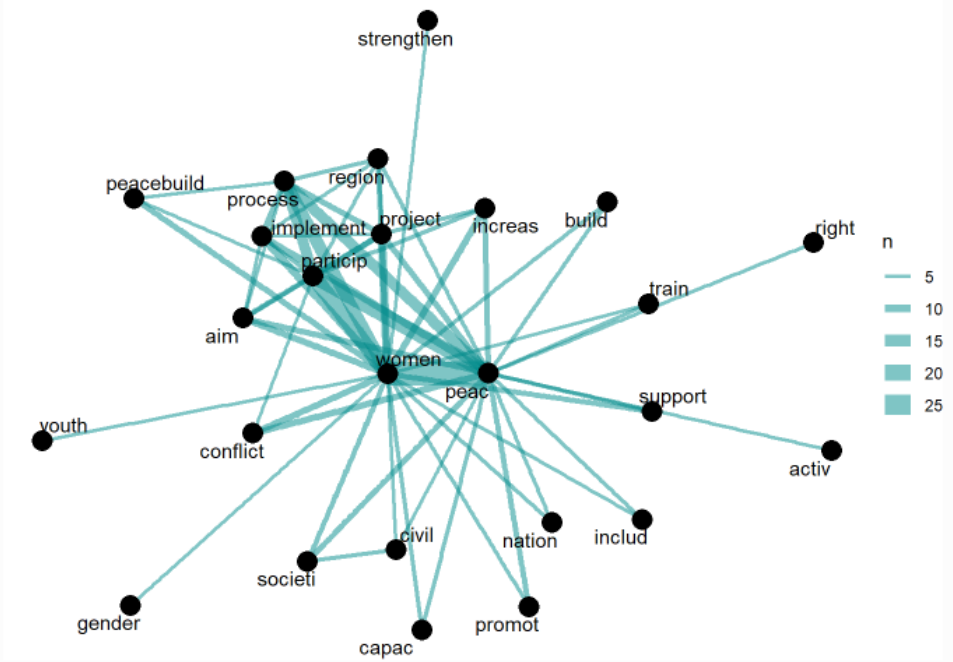


Textual descriptors in OECD DAC db

1	Agreement number	Agreement title	Impl Inst	Group of I Agreeer	Agreement Agr compl	Committer	Committer	Agreement	Agreement	Description of agreement	DAC Main	DAC Main	DAC Su	DAC				
30825	QZA-10/0953-77	FGM Awareness Creation and mobilization	HI - Habiba International	NGO Local	2011-2015	20110101		0	0	1080	192,6989			Female Genital Mutilation (FGM) awareness cre	130	130 - Pop	20	20 - F
30826	QZA-10/0926-21	Poverty eradication through child rights in	Plan International	NGO Inter	2011-2015	20110101		0	0	4568,471	815,1288			Given support, the program will empower coun	151	151 - Gove	30	30 - L
30827	FK13/108905AF	Peace Corps, FK Norway, personnel exchan	ANPPCAN - African Network for	NGO Local	2013-2013	20130101	20131231	0	0	30,658	5,21572			Promoting the Rghts of Children in Africa	151	151 - Gove	50	50 - L
30828	FK13/108905AF	Peace Corps, FK Norway, personnel exchan	ANPPCAN - African Network for	NGO Local	2013-2013	20130101	20131231	30,658	5,21572					Promoting the Rghts of Children in Africa	151	151 - Gove	50	50 - L
30829	MYS-13/0002	Navigating Women's Rights within the Conti	UNDP - UN Development Progi	Multilater	2013-2013	20130311		0	0	90	15,31133			Support to a seminar. The purpose of the progr	151	151 - Gove	70	70 - V
30830	MYS-13/0002	Navigating Women's Rights within the Conti	UNDP - UN Development Progi	Multilater	2013-2013	20130311		90	15,31133					Support to a seminar. The purpose of the progr	151	151 - Gove	70	70 - V
30831	NPL-13/0021	Strengthening Research, Education and Adv	Tribhuvan University	Public sec	2013-2020	20131210		0	0	19033,76	3238,135			The main objective of this project is to build hu	114	114 - Post	20	20 - F
30832	NPL-13/0021	Strengthening Research, Education and Adv	Tribhuvan University	Public sec	2013-2020	20131210		19033,76	3238,135					The main objective of this project is to build hu	114	114 - Post	20	20 - F
30833	RAF-12/0088	EAFF Programme funding 2012 – 2014	EAFF - Eastern African Farmers	NGO Local	2012-2014	20121214		0	0	8490	1460,042			The overall project goal is to contribute to stren	311	311 - Agric	94	94 - A
30834	SRB-13/0062	Roma Inclusion Office-Employment of Roma	Roma Inclusion Office	NGO Local	2013-2014	20131118		328,296	55,85165					The proposal offers clear benefits to the benefi	160	160 - Othe	20	20 - F
30835	SRB-13/0062	Roma Inclusion Office-Employment of Roma	Roma Inclusion Office	NGO Local	2013-2014	20131118		0	0	328,296	55,85165			The proposal offers clear benefits to the benefi	160	160 - Othe	20	20 - F
30836	QZA-10/0960-38	Africa. Education and Training	OATUU - Organisation of Africa	NGO Local	2011-2014	20110101		0	0	1420,18	253,3954			Organisational, Health, Safety and Environment	151	151 - Gove	60	60 - F
30837	QZA-13/0674	Partnership for Action on Green Economy P	UNEP - UN Environment Progr	Multilater	2013-2016	20131210		38330	6520,925					Programme support 2014: 7-year programme re	410	410 - Gen	10	10 - F
30838	QZA-13/0674	Partnership for Action on Green Economy P	UNEP - UN Environment Progr	Multilater	2013-2016	20131210		0	0	38330	6520,925			Programme support 2014: 7-year programme re	410	410 - Gen	10	10 - F
30839	RSA-12/0024	Cooperation project between the ANC and	Det norske Arbeiderparti	NGO Norv	2013-2014	20130307		1375	233,9231					Cooperation project between ANC and the Nor	151	151 - Gove	52	52 - L
30840	RSA-12/0024	Cooperation project between the ANC and	Det norske Arbeiderparti	NGO Norv	2013-2014	20130307		0	0	1375	233,9231			Cooperation project between ANC and the Nor	151	151 - Gove	52	52 - L
30841	PAK-10/0052	School Construction in Talgran Union Coun	Bulandi Welfare and Education	NGO Local	2010-2011	20101210		0	0	4610	762,6768			This grant shall be used to construct: (1) two gov	111	111 - Educ	20	20 - F
30842	UGA-11/0052	Strengthen Women Entrepreneurs phase2	Enterprise Uganda	Public sec	2013-2016	20131118		0	0	12000	2041,511			Strengthening Women Entrepreneurs project(S	321	321 - Indu	30	30 - S
30843	UGA-11/0052	Strengthen Women Entrepreneurs phase2	Enterprise Uganda	Public sec	2013-2016	20131118		12000	2041,511					Strengthening Women Entrepreneurs project(S	321	321 - Indu	30	30 - S
30844	NPL-13/0030	Governance Dialogue and Capacity Building	Royal Norwegian Embassy, Kat	Norwegia	2013-2015	20130912		1000	170,1259					This is an internal project of the Governance Tea	151	151 - Gove	12	12 - F
30845	QZA-12/0735-4	Women's Leadership Networks	Utdanningsforbundet	NGO Norv	2013-2015	20130101		2050	348,7581					Training/leadership programs and seminars to p	151	151 - Gove	70	70 - V
30846	QZA-12/0735-4	Women's Leadership Networks	Utdanningsforbundet	NGO Norv	2013-2015	20130101		0	0	2050	348,7581			Training/leadership programs and seminars to p	151	151 - Gove	70	70 - V
30847	GLO-09/982	Rapid Social Response	IBRD - International Bank for R	Multilater	2013-2014	20130614		70000	11908,81					The objectives of the programme are to safegua	160	160 - Othe	10	10 - S
30848	GLO-09/982	Rapid Social Response	IBRD - International Bank for R	Multilater	2013-2014	20130614		0	0	70000	11908,81			The objectives of the programme are to safegua	160	160 - Othe	10	10 - S
30849	UGA-13/0015	1300840 Strengthening Media in Post-Conf	Makerere University	Public sec	2013-2020	20131205		0	0	20600	3504,593			Makerere - HiOA - Nepal - South Sudan	114	114 - Post	20	20 - F
30850	UGA-13/0015	1300840 Strengthening Media in Post-Conf	Makerere University	Public sec	2013-2020	20131205		20600	3504,593					Makerere - HiOA - Nepal - South Sudan	114	114 - Post	20	20 - F
30851	ARM-09/018	Strengthening of the Sexual Assault Crisis C	WRC - Women's Resource Cen	NGO Local	2009-2012	20090924		0	0	2000	318,5525			Combat sexual violence against women and the	151	151 - Gove	70	70 - V
30852	QZA-11/0896-12	Regional initiatives of Latin American CSOs	Norsk Folkehjelp	NGO Norv	2012-2015	20120101		0	0	3803,528	654,1003			Partners to participate in regional networks,	151	151 - Gove	50	50 - L
30853	QZA-10/0960-1	Africa. Organisational development	ZCTU - Zimbabwe Congress of	NGO Local	2011-2014	20110101		0	0	5297,4	945,1879			General support to Zimbabwe Congress of Trade	151	151 - Gove	60	60 - F
30854	QZA-12/0763-127	NMA- Community Development Project, Fc	NMA - Misjonsalliansen	NGO Norv	2013-2015	20130101		0	0	3020,556	513,8748			Community development of marginal areas of G	160	160 - Othe	61	61 - C
30855	QZA-12/0763-127	NMA- Community Development Project, Fc	NMA - Misjonsalliansen	NGO Norv	2013-2015	20130101		0	0	3020,556	513,8748			Community development of marginal areas of G	160	160 - Othe	61	61 - C
30856	QZA-12/0763-127	NMA- Community Development Project, Fc	NMA - Misjonsalliansen	NGO Norv	2013-2015	20130101		3020,556	513,8748					Community development of marginal areas of G	160	160 - Othe	61	61 - C
30857	INS-13/0038	Triangular cooperation on Exchange Afghan	NAC - Afghanistankomiteen	NGO Norv	2013-2017	20131204		3954	672,6778					The Goal is to increase access to quality educati	111	111 - Educ	30	30 - T
30858	INS-13/0038	Triangular cooperation on Exchange Afghan	NAC - Afghanistankomiteen	NGO Norv	2013-2017	20131204		0	0	3954	672,6778			The Goal is to increase access to quality educati	111	111 - Educ	30	30 - T
30859	QZA-10/0953-630	Women empowerment through literacy ed	ADA - Afghanistan Developme	NGO Local	2013-2015	20130101		0	0	1083,084	184,2606			Providing 9 months basic literacy education to r	112	112 - Basic	20	20 - F

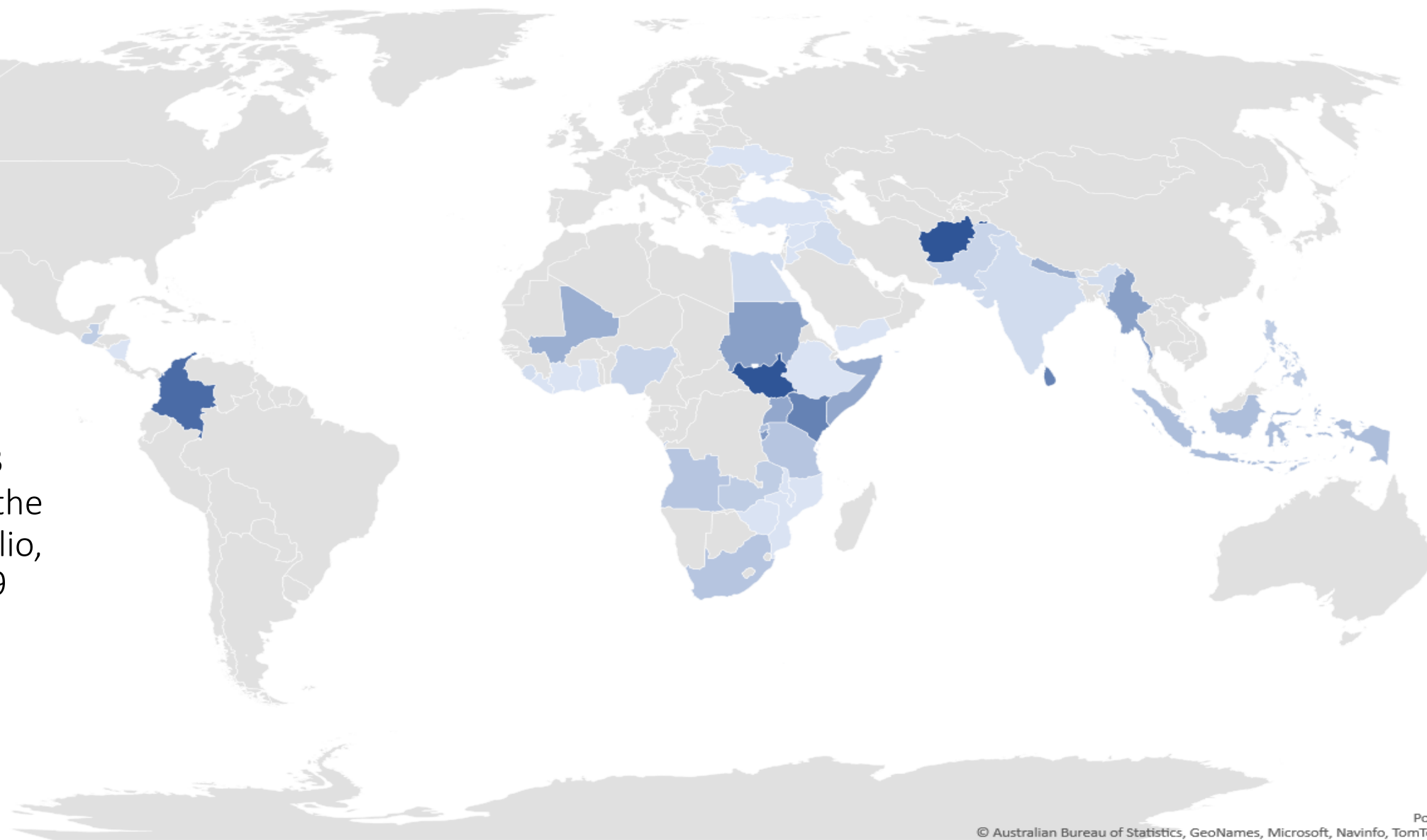
Machine learning

- R
- Supervised learning
- Steps after cleaning and preparing the dataset of WPS interventions, 2000 – 2019:
 1. Selecting a portion (<300) of the total observations (training set);
 2. Manual coding of the training set, looking at “Description of Agreement”. Is it a “women’s participation in peace” intervention? (1/0)
 3. Tokenisation and further cleaning (stopwords, stemming, etc.)
 4. Train the model (exploratory analysis, evaluation)
 5. Apply the model and get projects classified



Results

- At least 333 projects in the WPS portfolio, 2000 - 2019



% all projects across countries



Final thoughts

- Output quality - input quality.
- Portfolio analyses and text mining:
 - A wide range of opportunities.
 - Supervised classification as a cost-effective method.
- Evaluation skills, thematic expertise and data science competence.

Follow us

- Newsletter: www.norad.no/evalnews
- Twitter: @noradevaluation

**Session 3: Using Machine Learning
for Analysing Human Rights
Mainstreaming in Development
Project Documents**

Lea Smidt, Evaluator, DEval

Using Machine Learning for Analysing Human Rights Mainstreaming

EES Webinar

Lea Smidt | Deval

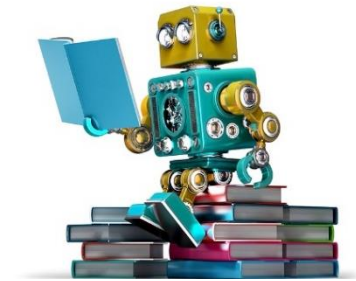
15.04.2021

Page 2

Machine Learning and Text

Text analysis allows...

- ...**extracting information** from text
- ...establishing a **relationships** between **words** and **variables** ('classes')



Text analysis and machine learning

- **Computer 'learns'** to decide whether text belongs to a **class** or not based on words
- Two 'learning' approaches

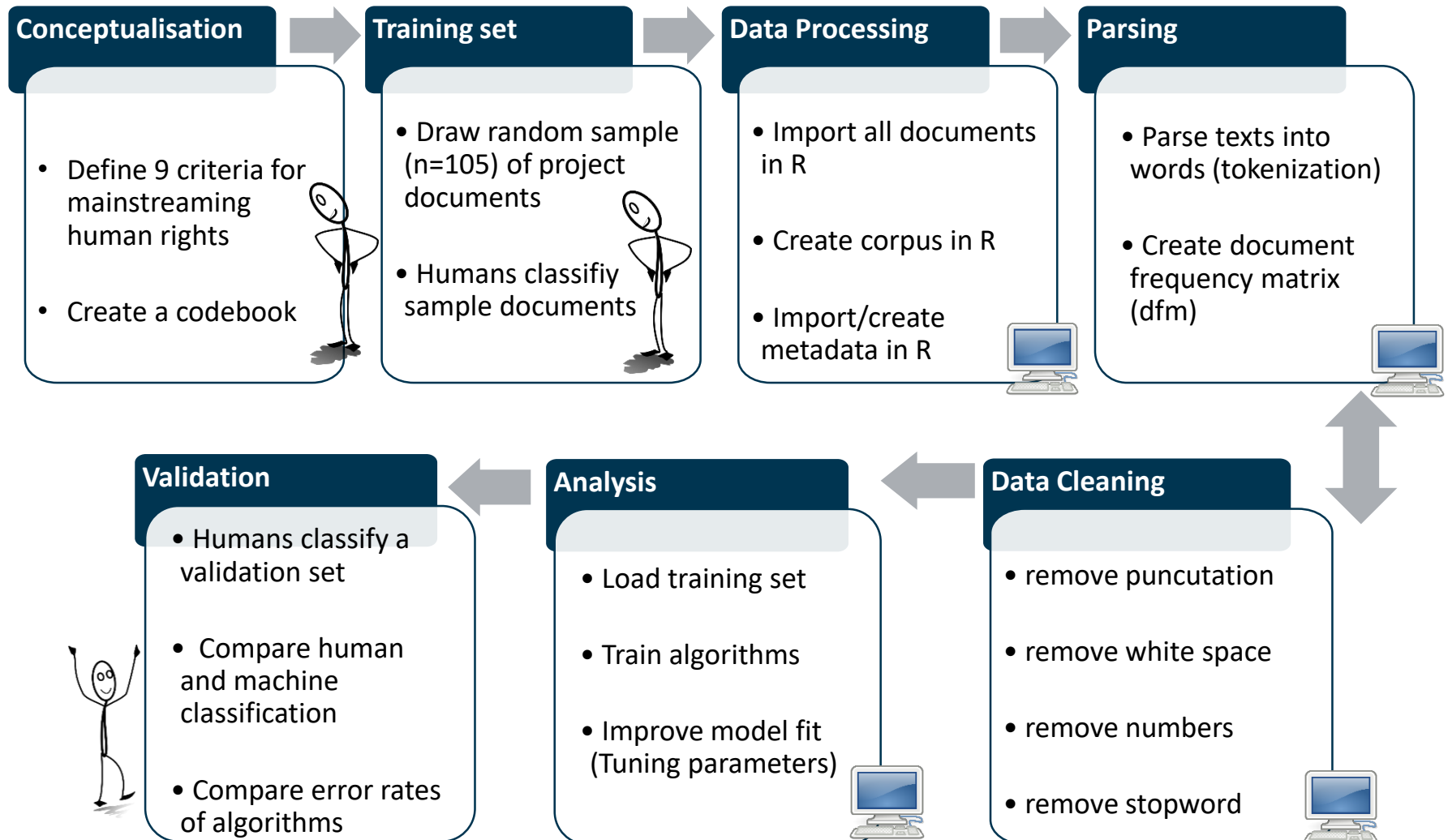
Supervised learning

- Evaluator classifies documents (training data)
- Known classes
- Computer 'learns' through training, feedback & repetition

Unsupervised learning

- Computer determines word clusters
- Unknown classes
- Computer 'learns' through repetition

Work Flow: Supervised Machine Learning



Challenges, Risks and Safeguards

Lack of information on the **data-generating process**

- What do the documents stand for?
- Who created the document in which context?

Biased results due to...

- ...different **types of documents** and **coding units**
- ...an unrepresentative **training data set**
- ...a reproduction of **human errors**

Validation

- Underlying assumption: Humans are always right (...actually they are not!)
- Compare best-performing algorithm to intercoder-reliability of humans

➔ **Be transparent about the limitations of your data and analyses**

Thank you very much for your
attention

Further Reading

Cingranelli, David, und David L. Richards (2014). „The CIRI Human Rights Data Project Coding Manual“. Version 5.20.14.
<http://www.humanrightsdata.com/p/data-documentation.html>.

Evangelopoulos, Nicholas, und Lucian Visinescu (2012). „Text-Mining the Voice of the People“. *Communications of the ACM* 55(2): 62.
<https://doi.org/10.1145/2076450.2076467>.

Früh, Werner (2017). *Inhaltsanalyse: Theorie und Praxis*. UTB Medien- und Kommunikationswissenschaft, Psychologie, Soziologie 2501. Konstanz: UVK Verlagsgesellschaft mbH.

Greene, Kevin T., Baekwan Park, und Michael Colaresi (2019). „Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects“. *Political Analysis* 27(2): 223–30.
<https://doi.org/10.1017/pan.2018.11>.

Grimmer, Justin, und Brandon M. Stewart (2013). „Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts“. *Political Analysis* 21(3): 267–97. <https://doi.org/10.1093/pan/mps028>.

Hopkins, Daniel J., und Gary King (2010). „A Method of Automated Nonparametric Content Analysis for Social Science“. *American Journal of Political Science* 54(1): 229–47. <https://doi.org/10.1111/j.1540-5907.2009.00428.x>.

James, Gareth, Daniela Witten, Trevor Hastie, und Robert Tibshirani (2017). *An Introduction to Statistical Learning: With Applications in R*. Corrected at 8th printing. Springer Texts in Statistics. New York Heidelberg Dordrecht London: Springer.

Minhas, Shahryar, Jay Ulfelder, und Michael D Ward (2015). „Mining Texts to Efficiently Generate Global Data on Political Regime Types“. *Research & Politics* 2(3). <https://doi.org/10.1177/2053168015589217>.

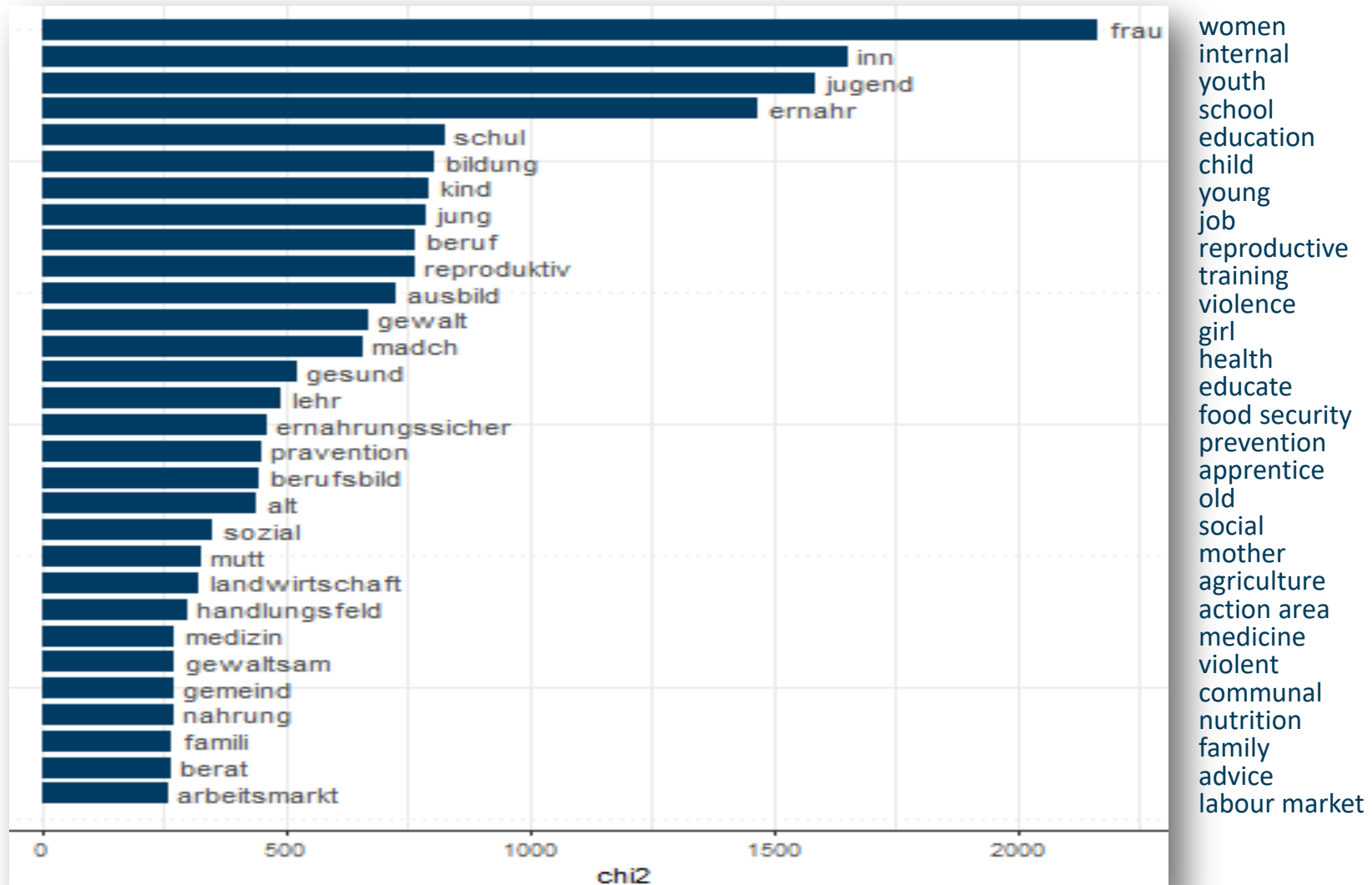
Sanditya Hardaya, I. B. N., Arian Dhini, und Isti Surjandari (2017). „Application of text mining for classification of community complaints and proposals“. In 2017 3rd International Conference on Science in Information Technology (ICSITech), 144–49. Bandung: IEEE. <https://doi.org/10.1109/ICSITech.2017.8257100>.

Slapin, Jonathan B., und Sven-Oliver Proksch (2014). „Words as Data“. In Shane Martin, Thomas Saalfeld, und Kaare W. Strøm (2014). *The Oxford Handbook of Legislative Studies*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199653010.013.0033>.

BACK UP

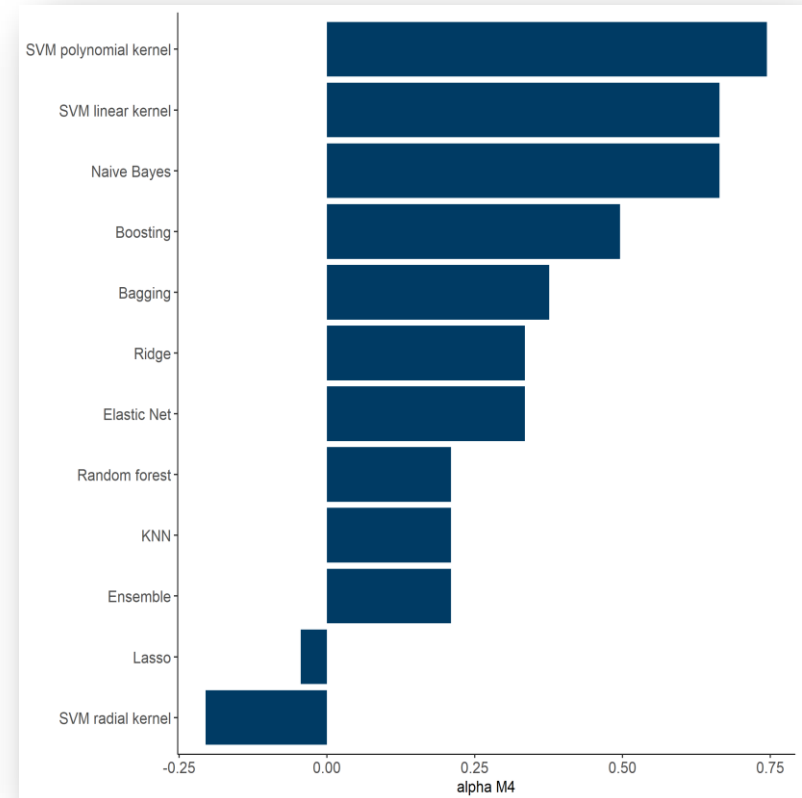
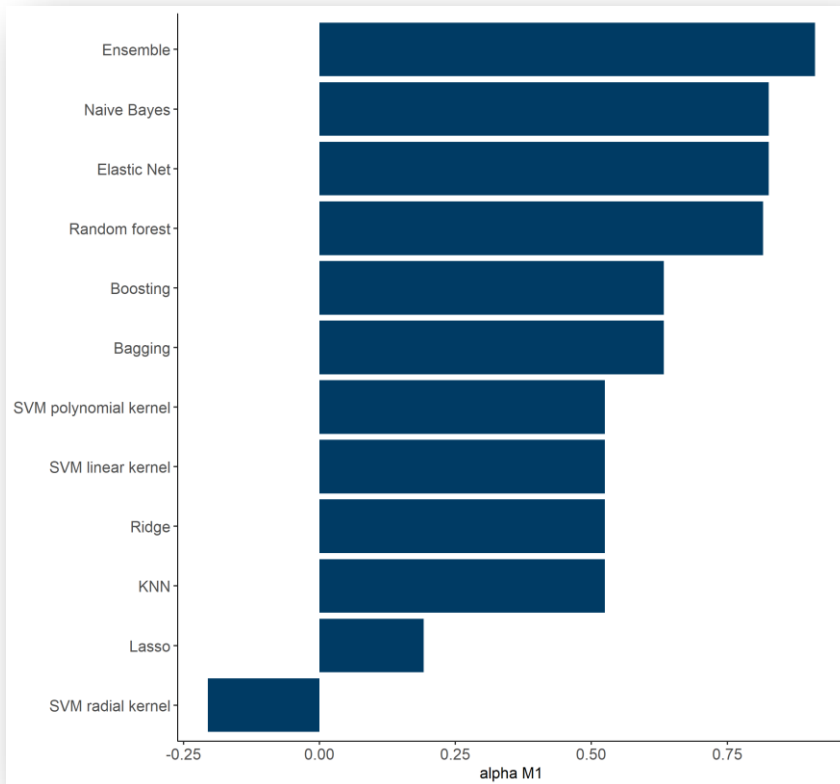
Results II: Content of HR Mainstreaming

Criterion 4: Empowerment of Marginalised Groups



Performance and reliability

Advise: Compare best-performing algorithm to intercoder-reliability of human coders



Class: Strengthen state actors' accountability and transparency

Class: Empowerment

Discussion with the Audience

Facilitated by Estelle Raimondo, EES
Board Member

Closing Remarks

Alena Lappo, EES Board Member