Measuring the quality of large-scale automated classification systems applied to online job advertisement data

Webinar ESSNET WIN

Alexander Kowarik, Johannes Gussenbauer, Magdalena Six

05.03.2024

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

Content

Input Quality:

Quality of the Sources
Presentation by Magdalena Six

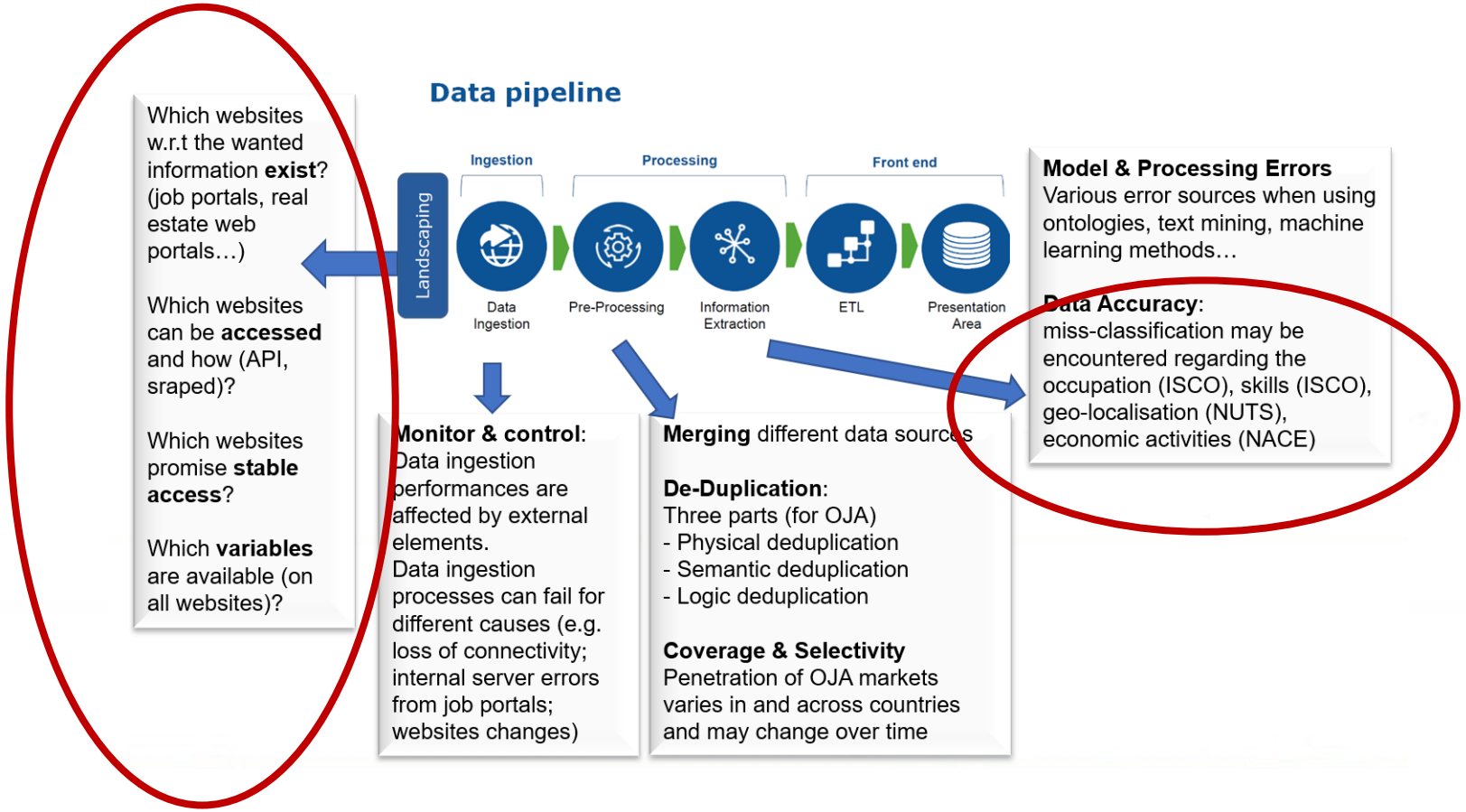
Process / Output Quality

Quality of the Classifications

Presentations by Alexander Kowarik and Johannes Gussenbauer

Funded by the European Union

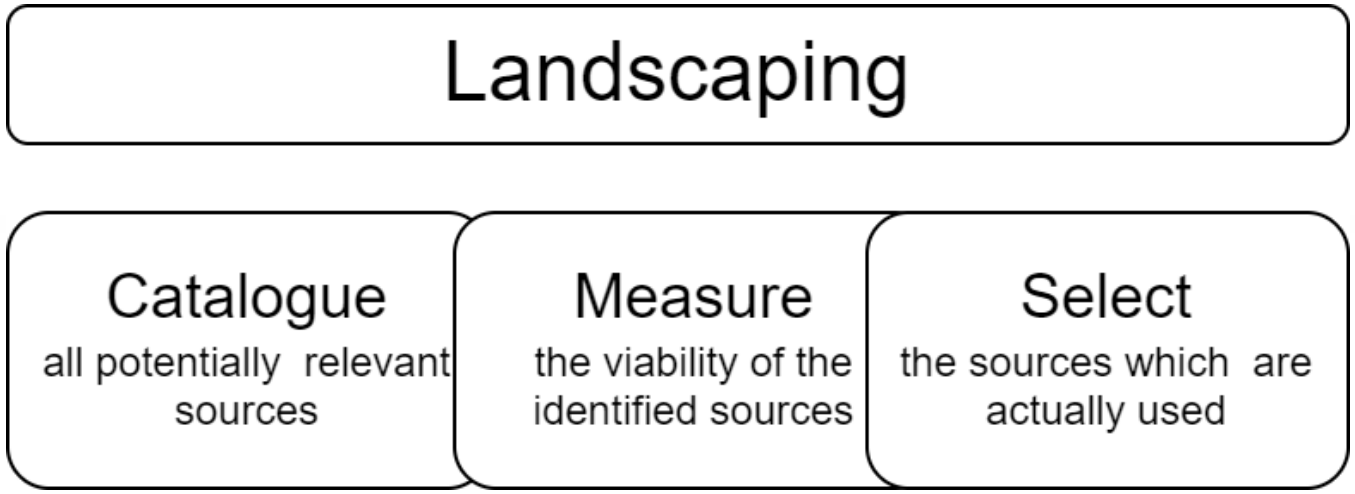
Content


Which websites w.r.t the wanted information **exist**? (job portals, real estate web portals…)
Which websites can be **accessed** and how (API, sraped)?
Which websites promise **stable access**?
Which **variables** are available (on all websites)?
Data pipeline
Landscaping
Ingestion
Processing
Front end
Data Ingestion
Pre-Processing
Information Extraction
ETL
Presentation Area
Monitor & control: Data ingestion performances are affected by external elements. Data ingestion processes can fail for different causes (e.g. loss of connectivity; internal server errors from job portals; websites changes)
Merging different data sources
De-Duplication: Three parts (for OJA)
- Physical deduplication
- Semantic deduplication
- Logic deduplication
Coverage & Selectivity Penetration of OJA markets varies in and across countries and may change over time
Model & Processing Errors Various error sources when using ontologies, text mining, machine learning methods…
Data Accuracy: miss-classification may be encountered regarding the occupation (ISCO), skills (ISCO), geo-localisation (NUTS), economic activities (NACE)

Slido Question

Landscaping

Definition: **Landscaping** comprises all process steps necessary to **catalogue** all relevant sources for a specific topic of interest, to **measure** the viability of the catalogued sources and to **select** the sources, which are actually used, based on the measured criteria.



Landscaping

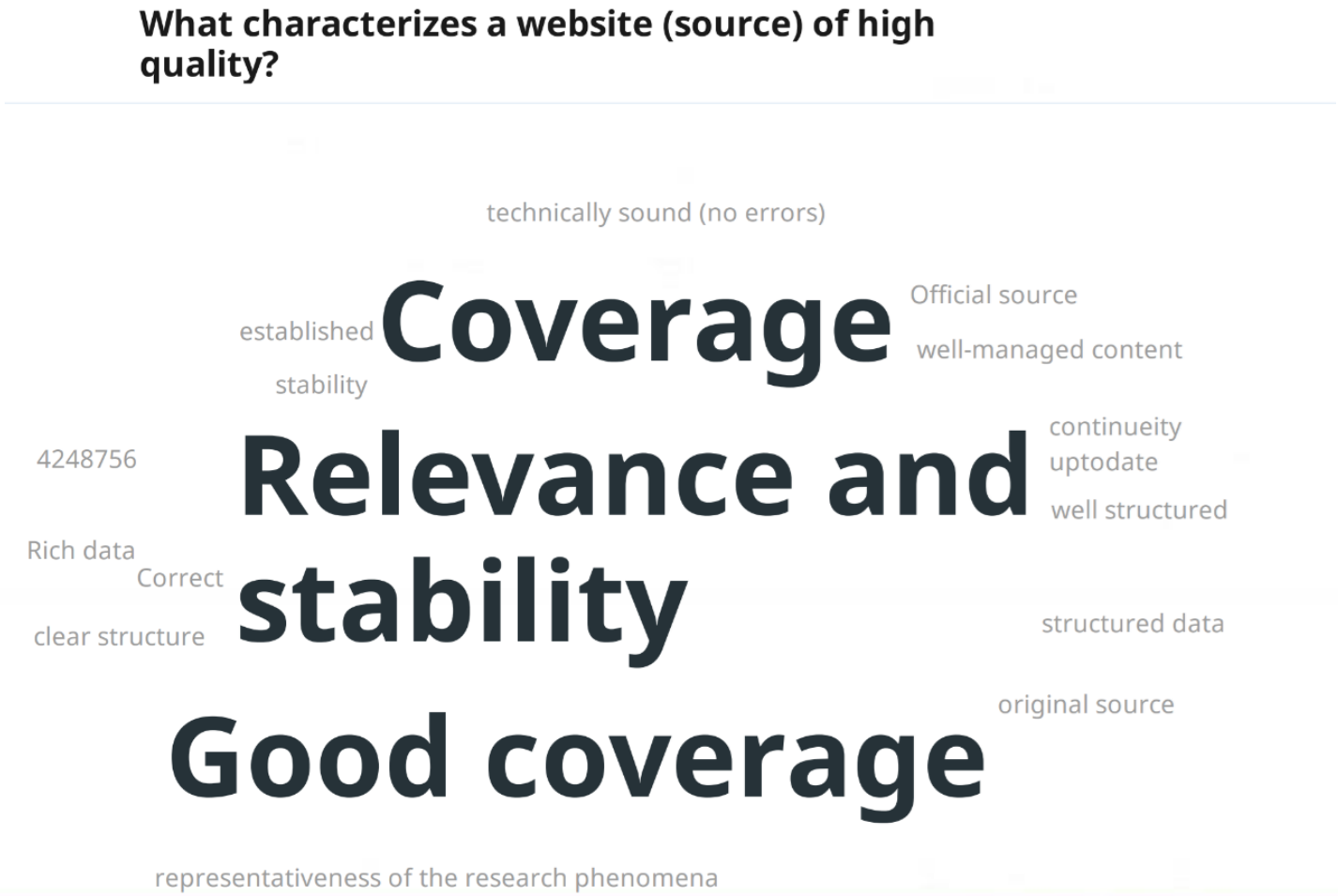
Important Question:

Does your topic of interest require to

- find **all websites?**
  Examples: online based enterprise characteristics, typification of enterprises according to “green industry”
- find **list of representatives** which fulfil certain criteria?
  Examples: OJA portals, prices of specific goods, real estate websites, booking platforms…

More about landscaping in Deliverable 4.3, partial draft already accessible here:
https://webgate.ec.europa.eu/fpfis/wikis/display/WIN/Deliverables+and+Milestones

Slido Question



Input quality of OJA data

Landscaping of sources for OJA data was done **centrally by Eurostat.**

NSIs are more in the role of **users** of the collected and processed data.

WP4 developed **quality indicators** for the assessment of the selected OJA sources

- **Relevance** of the source
- **Stability** w.r.t the **existence** of the sources
- **Stability** w.r.t the **popularity** of the sources

Why are dynamic sources a problem?

Goal:
Capture the **dynamics of the labour market** with the help of scraped OJA
-> need some form of indicator aggregated over several sources / index

Problem:
Scraped OJAs might (additionally) **capture the dynamics of the sources** (concept drift)

How to measure this?

Funded by the European Union

Quality Indicators about Relevance

Indicator 1a

If your NSI scrapes OJA data itself, **compare the included sources** from your own scraping processes with the included sources on the Web Intelligence Platform (WIP).

Indicator 1b

If your NSI does not scrape**, consult the labour market experts** in your NSI and ask them to **name the x most important job portals** in your country and compare this list with the sources on the WIP for your country

Stability of existence of the sources

Indicator 2

Calculate the **number of sources over time**.

Indicator 3a and 3b

Determine if it is **always the same sources** in the course of the time span considered

Determine if the most **important sources at several points of time** are **present over the whole time span.**

Funded by the European Union

Stability of the popularity of the sources

Example: Increase of number of scraped OJAs of one source.
- due to increase of open positions in labour market?
- due to increase of popularity of the source?

Indicator 4 and 5

Calculate the **ranking of the most important sources** w.r.t the OJA volume and observe this ranking over the course of time

Plot the number of OJA per source over time and **check if the dynamics of the individual time series per source are similar**

Funded by the European Union

Quality assessment of the sources - Results

- Implementation of the quality indicators with the help of Rmarkdown script
- Executed centrally in Rstudio on WIP for each of the participating countries
- Standardized reports for each country available on WIN Confluence

https://webgate.ec.europa.eu/fpfis/wikis/display/WIN/Quality+assessment+of+OJA+sources

- Rmarkdown code available at GitHub:

https://git.fpfis.tech.ec.europa.eu/estat/wihp/analysis/oja_sample_annotation/-/tree/develop/quality%20indicators%20OJA?ref_type=heads

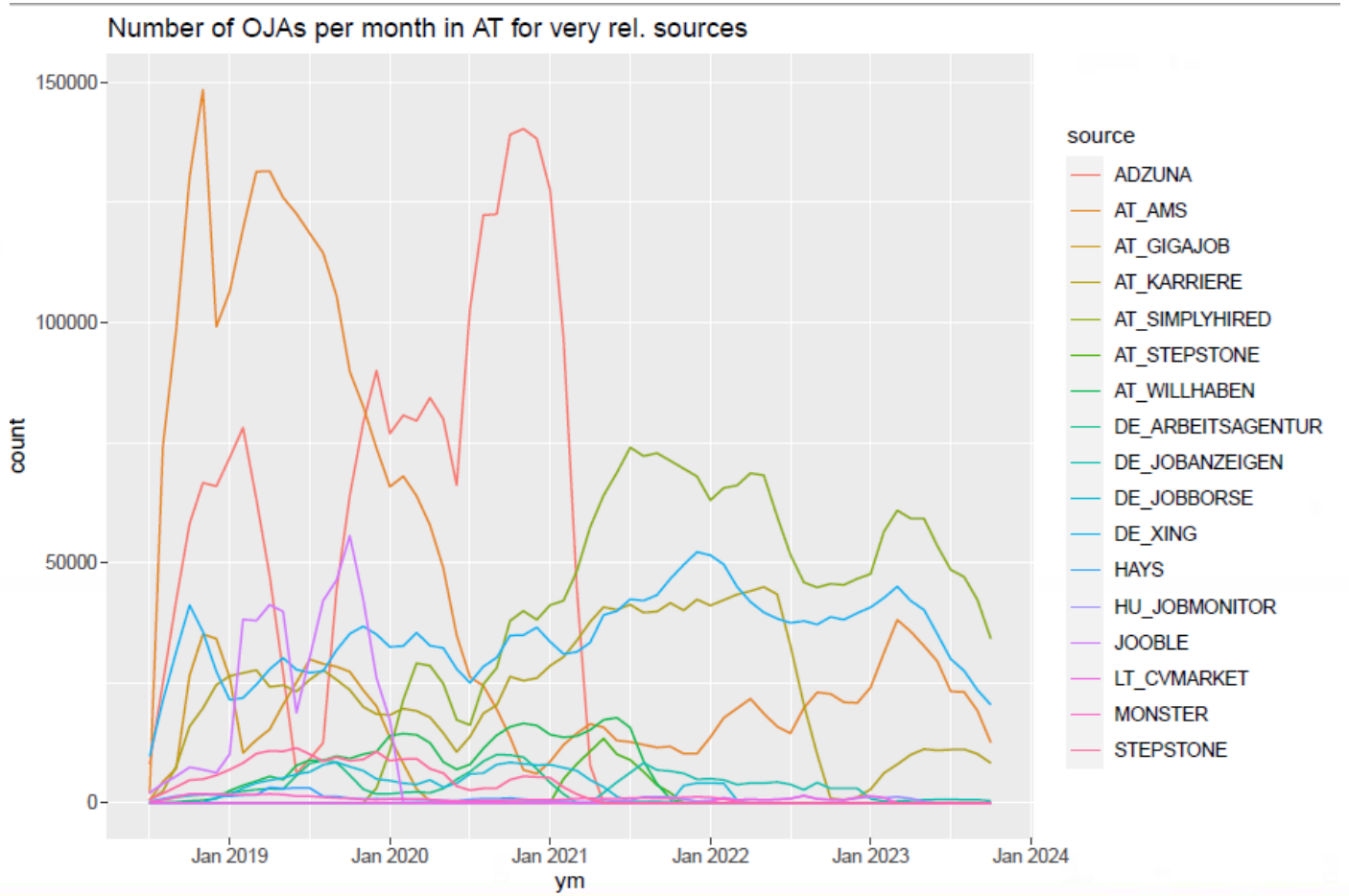
The number of very relevant sources (>5000 OJA) in different countries across the years.

Year	AT	BG	DE	FI	FR	IT	NL	PL	PT	RO
2018	7	1	27	2	24	16	9	7	2	2
2019	13	5	31	6	25	23	15	11	4	11
2020	9	4	28	4	22	17	12	8	10	7
2021	8	4	28	4	25	21	15	8	14	8
2022	7	5	26	3	31	19	14	11	16	9
2023	4	4	21	3	26	17	13	11	12	6

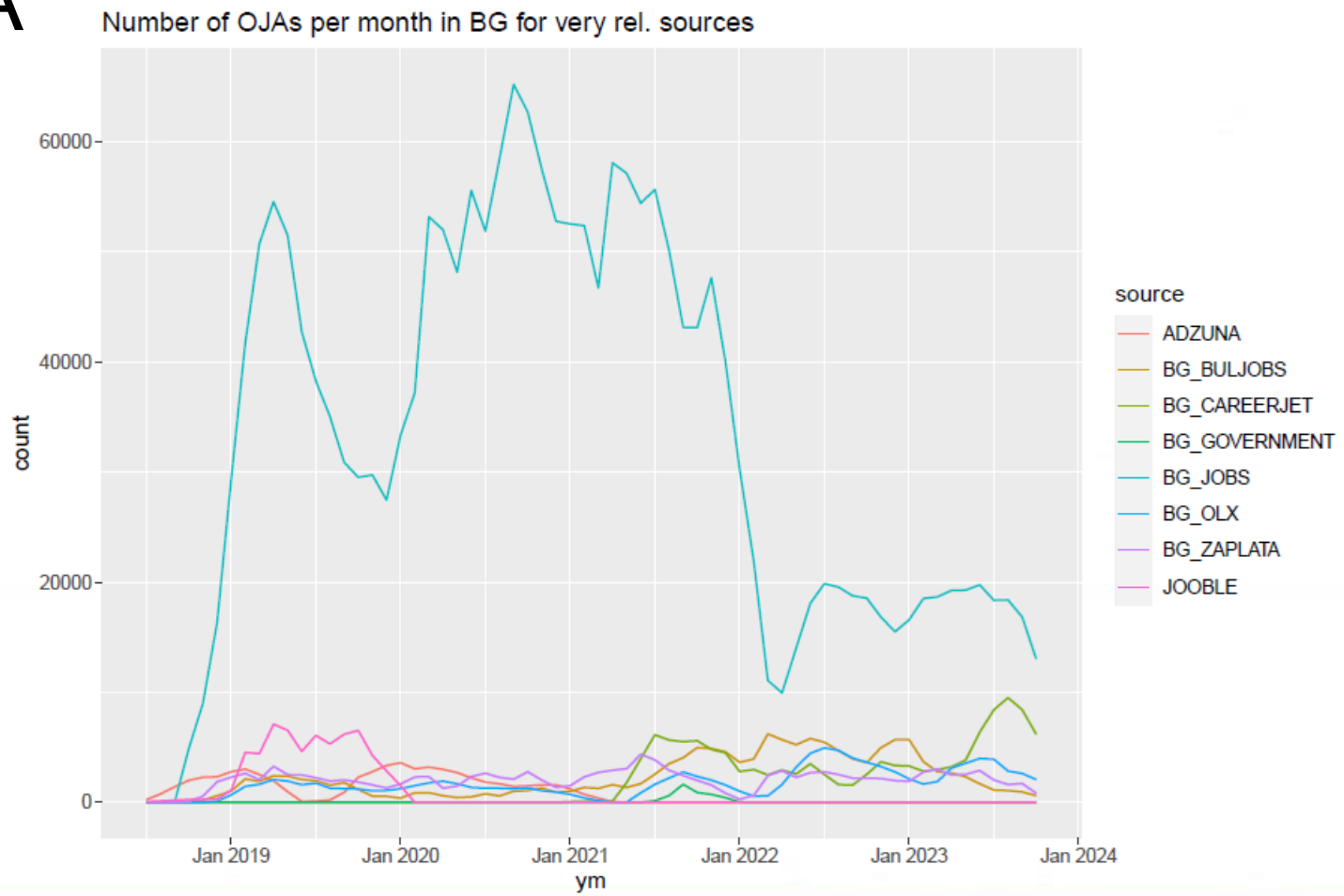
Results – Ranking of most popular sources, example NETHERLANDS

rank	2018	2019	2020	2021	2022	2023
1	108	108	108	781	781	781
2	578	410	642	642	465	992
3	560	465	427	427	452	452
4	465	578	452	452	468	468
5	427	468	468	550	992	465

Results – Dynamics of sources, example AUSTRIA



Results – Dynamics of sources, example BULGARIA



Measuring the quality of large-scale automated classification systems applied to online job advertisement data

Webinar ESSNET WIN

Alexander Kowarik, Johannes Gussenbauer, Magdalena Six

05.03.2024

Trusted Smart Statistics – Web Intelligence Network

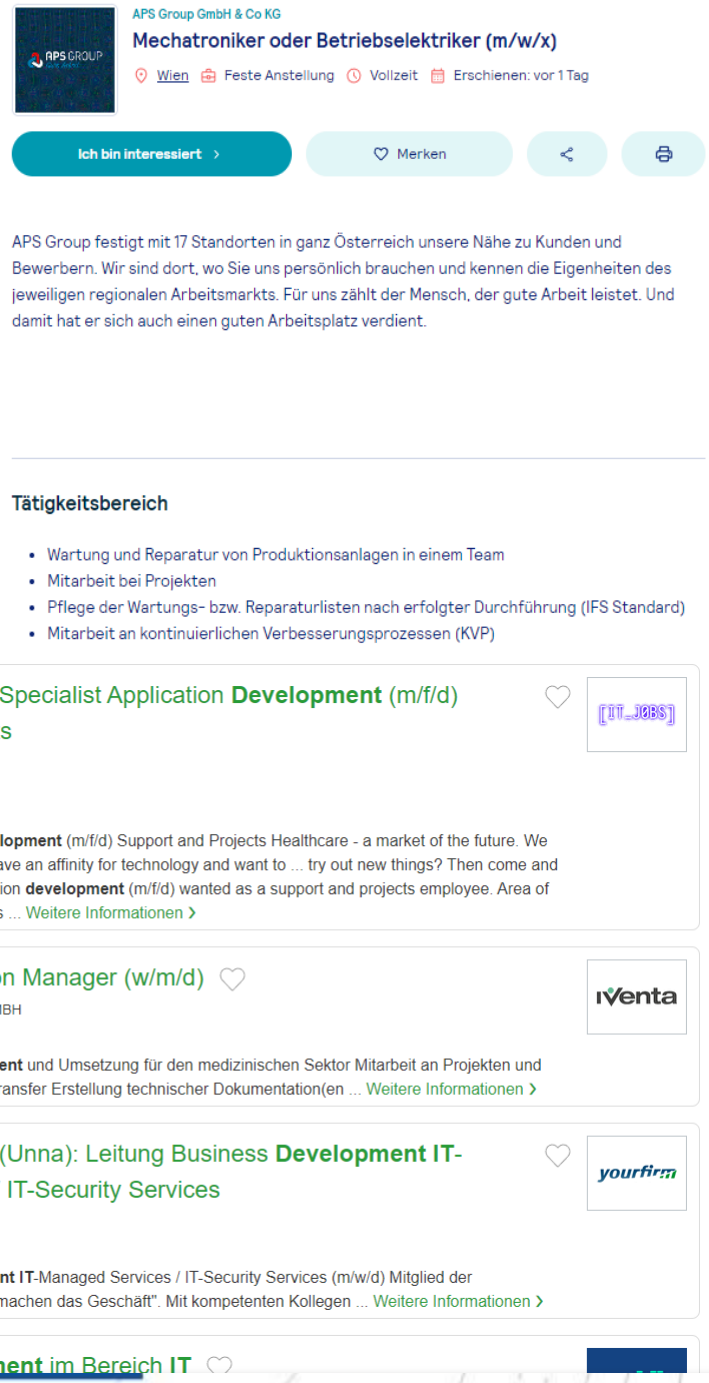
Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

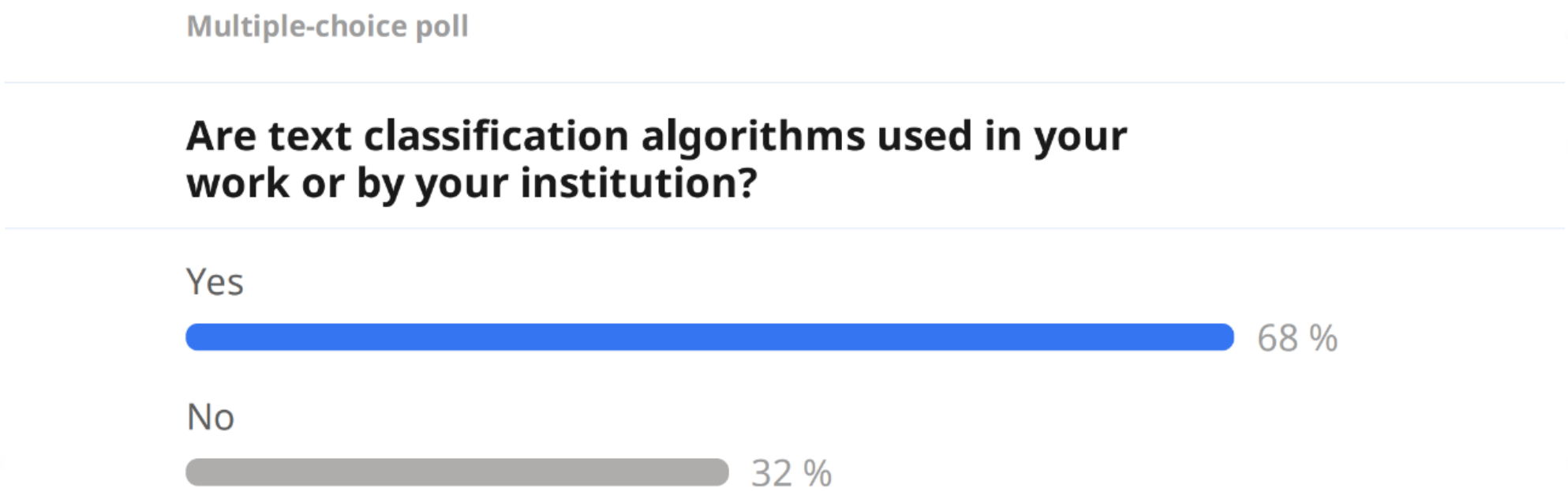
Classification is what makes job ads useful for statistics

- The web scraped Online Job Advertisement data needs to be classified in order to use it in production of statistics.
- The classification of the data is done automatically by using machine learning techniques.
- The accuracy of the classification algorithm needs to be assessed carefully.
- This can be done by annotation exercises, where a sample is drawn from the classified data and the classification is then checked by humans.



Web Intelligence Network

Slido Question



Funded by the European Union

Current classification methodology

Funded by the European Union

Classified dimensions

- Occupation: ISCO up to 4th digit
- Economic Activity of the company : NACE up to 2nd digit
- Minimal education required: ISCED 8 levels
- Region where the job is to be performed: NUTS up to LAU1 regions
- Number of hours: full time / part time

- Type of contract: Unlimited, limited, self employed, …
- Salary per year in salary brackets
- Years of working experience required
- …

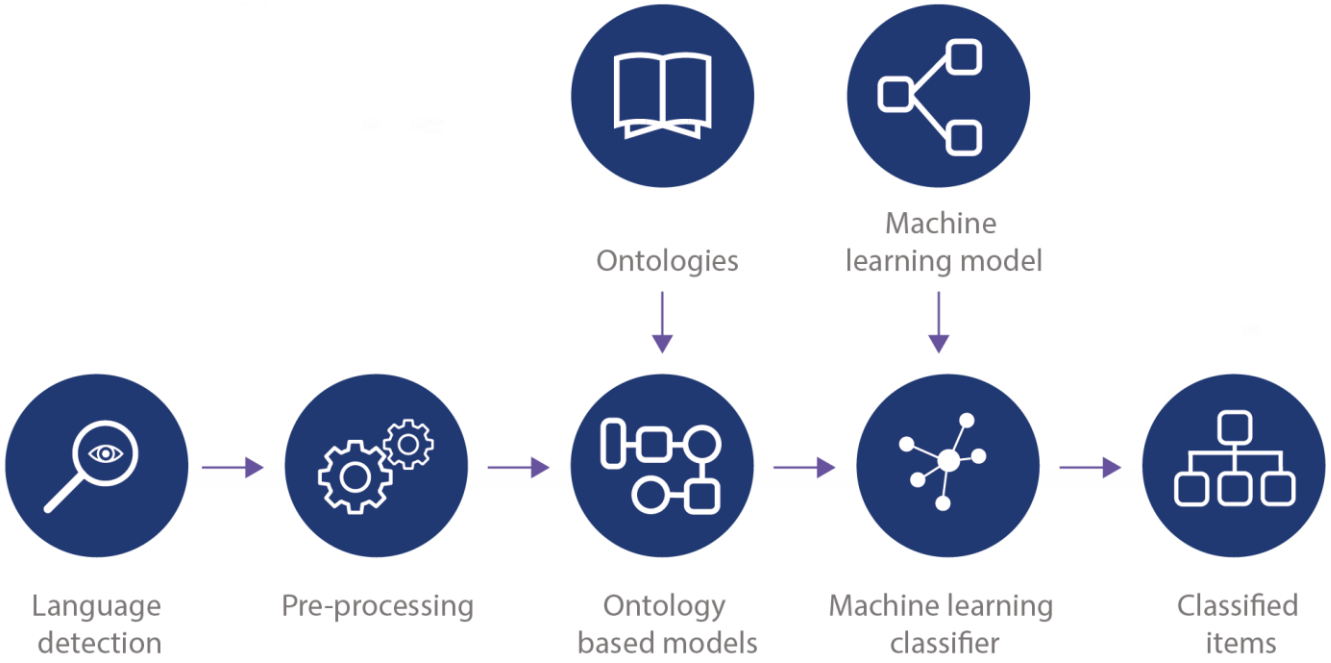
Funded by the European Union

Data Validation Process

- Set of validation rules for data processing
  - Consistency with Eurostat's official code lists
  - Consistency within hierarchical classifications
  - The distribution of ads within categories of a classification should be reasonably stable over time
  - categories should not be too unbalanced
- Constant validation cycle to mainly investigate sudden changes in for the classification system
- These rules **do not** address the performance of the classification itself.

Funded by the European Union

Setup for classification pipeline



Classification methods

- Language detection + pre-processing
- Ontology-Based model
  - String matching (exact and similar) between text from job ad and onotology
  - Job title vs job description
  - Stemming and lemmetization
- Machine learning classifier
  - Only used if ontology based model does not achieve predictions
- In most of the cases a label is extracted using the ontology-based model

Funded by the European Union

Setup annotation exercise

Web Intelligence
Network

Funded by
the European Union

Quality of classification can be assessd by manual inspection

- Annotation (= Labelling, manually classifying) of job ads according to a classification
- 1st – mid 2022 -> only occupation(ISCO), but detailed
- 2nd – end 2023 -> occupation, economic activity, education, location and working time, but only on 1st level
- High cost for manual inspection -> optimized sample and efficient tool

Funded by the European Union

Doccano as annotation tool

Web Intelligence Network

Funded by the European Union

Doccano as annotation tool II

- Set of labels to be assigned

Doccano as annotation tool III

- Input for annotation

Title: Agile Coach (m/w/d) /// Job description: Agile Coach (m/w/d) Zur Gesamtübersicht Merken Baden-Württemberg | Freiberuflich für ein Projekt Referenznummer 620116/1 Startdatum sofort Projektdauer 3 MM Jetzt bewerben Meine Aufgaben Scaled Einführung Aufbau Operation Security Center Meine Qualifikationen Scaled Agile Framework Erfahrung IT- Security Kenntnisse von Vorteil Wünschenswert: SoC / SIEM Sprache: Deutsch und Englisch Meine Vorteile Dynamisches Team Remote abbildbar Großkonzern Über Hays Der

- Values predicted by the algorithm

Key	Value
oja_id	1001467521
education	ED7 Master's or equivalent level
occupation1d	OC1 Managers
economic_activity1d	D Electricity, gas, steam and air c
working_time	NA Not indicated (missing)
nuts2	AT12 Niederösterreich

Doccano as annotation tool IV

Give instructions for annotators:

- What is the aim of the annotation exercise?
- What kind of data needs to be annotated?
- How to access the tool?
- What do the labels mean?
- When is a label correct/incorrect/missing?

	Correct	Incorrect	Wrong language	Not a job ad	Job description missing	Not indicated (missing)	Different Country
Education	Variable Education labelled correctly	Variable Education labelled incorrectly	-	-	-	Labelling Education is not possible given title and description	-
occupation1d	Variable occupation1d labelled correctly	Variable occupation1d labelled incorrectly	Job ad is not in official language	Shown text is not a job ad	Job description is not available	Labelling occupation1d is not possible given title and description	-
economic_activity1d	Variable economic_activity1d labelled correctly	Variable economic_activity1d labelled incorrectly	-	-	-	Labelling economic_activity1d is not possible given title and description	-
working_time	Variable working_time labelled correctly	Variable working_time labelled incorrectly	-	-	-	Labelling working_time is not possible given title and description	-
nuts2/nuts1	Variable nuts2/nuts1 labelled correctly	Variable nuts2/nuts1 labelled incorrectly	-	-	-	Labelling nuts2/nuts1 is not possible given title and description	Location of job is in a different country

Slido Question

Have you or your institute used annotation services (Mechanical Turk, ...) or tools (Doccano, ...) and which one?

The sample should well represent the different values of a classification

- 1st round -> stratified sample according to occupation

- 2nd round -> several target variables, optimal design?
  - When dealing with many classification variables and possibly many outcome values for each variable the number of annotated records needed to derive a high degree of accuracy can quickly reach tens of thousands.

Funded by the European Union

Sample design – annotation (2nd round)

- Optimize marginal distribution of each variable instead of joined

Dimension	ED 1	ED 2	ED 3	ED 4	ED 5	ED 6	ED 7	ED 8	Not indicated
Education level	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33

Dimension	FT	PT	Not indicated
Working time	100	100	100

Dimension	OC 1	OC 2	OC 3	OC 4	OC 5	OC 6	OC 7	OC 8	OC 9
Occupation 1st Digits	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33

Dimension	A	C	D	E	...	Q	R	S	T
Economic Activity (Sections)	15.79	15.79	15.79	15.79	...	15.79	15.79	15.79	15.79

Dimension	AT 11	AT 12	AT 13	AT 21	AT 22	AT 31	AT 32	AT 33	AT 34	Not indicated
NUTS2	30	30	30	30	30	30	30	30	30	30

Heuristic algorithm for optimal design

Implemented in R package simPop function calibPop: https://CRAN.R-project.org/package=simPop

1. Randomly initialize a sample of size n and set starting temperature T
2. Compare the margins resulting from the sample $\{\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_M\}$ to the target margins $\mathbf{t}_1, \ldots, \mathbf{t}_M$ and calculate the initial value of the objective function $Obj_0 = f(\mathbf{t}_1, \ldots, \mathbf{t}_M, \hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_M)$.
3. Randomly add and discard some records from the sample
   - Sample with probability according to over- or under-representation in current target margins $\{\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_M\}$
4. Re-calculate $\{\hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_M\}$. If the difference between sample and target margins is small enough → stop, otherwise go to 4.
5. Check if current solution has become better or worse than previous one $Obj_s < Obj_{s-1}$
   - Accepts worse solution with a probability of $\exp\left(-\frac{Obj_s - Obj_{s-1}}{T}\right)$
6. Cooldown T by fixed factor.
7. Terminate if maximum number of iterations has been reached otherwise go to step 2

Results of annotation

Web Intelligence Network

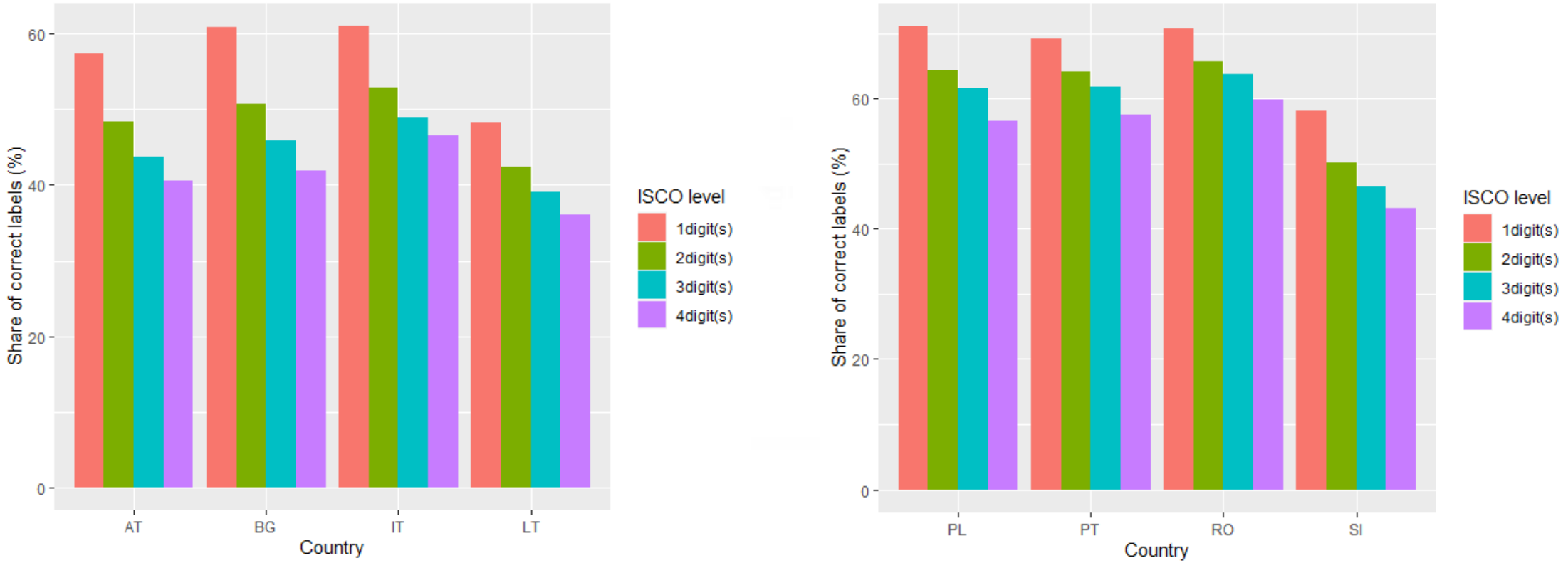
Funded by the European Union

Quality is increasing for higher level of aggregation

Classification	1digit(s)	2digit(s)	3digit(s)	4digit(s)
Correct	62.04	54.83	51.40	47.81
Incorrect	30.77	37.59	40.52	43.53
Incorrect - Missing label	1.16	1.26	1.47	1.73
Impossible to classify	0.74	0.74	0.89	1.61
No reference to occupation/Job description missing	4.40	4.69	4.83	4.80
Not a job ad	1.28	1.28	1.28	1.28
Wrong language	1.92	1.64	1.75	1.75

Table 6: Average percentage shares of each class assigned by the annotators in all countries.

Results look similar for all countries.



Percentage shares of correct labels for different ISCO levels for Poland, Portugal, Romania and Slovenia.

7 NSIs annotated 2442 OJAs according to 5 dimensions

- 7 countries: Austria, Bulgaria, Finland, France, Italy, Poland and Slovenia
- Total of 2442 OJAs are annotated
- 5 dimensions labelled:
  - Economic activity
  - Education
  - Occupation
  - Location
  - Working time

Funded by the European Union

The annotator/expert is not that important?!

- All ads of PL were labelled twice, independently by two different experts

- We can compare how similar their judgement was.
  - Percentage of ads where both annotators marked „correct“, „incorrect“ the same way:
  - Economic activity: 94 %
  - Education: 94 %
  - Occupation: 87 %
  - Location: 93 %
  - Working time: 88 %

Funded by the European Union

Different weights allow for different ”perspectives” on the results

- Unweighted = equally weighted: All classes of a classification are equally important
- Weighted: Classes have weight according to their frequency per country
- First number = unweighted
- Second number = weighted

Funded by the European Union

Results 2nd annotation – economic activity, working time

economic_activity 1d_correct	AT	BG	FI	FR	IT	PL	SI
NA	0	5.1 / 4.1	0.9 / 0.8	3 / 3	1.6 / 1.9	0.2 / 0.5	14.3 / 10.7
correct	31.8 / 22.2	29.6 / 26.3	30.6 / 27.6	33 / 29.1	27.7 / 32.8	20.9 / 21.4	31.6 / 22.4
incorrect	68.2 / 77.8	65.3 / 69.6	68.5 / 71.6	64 / 67.9	70.6 / 65.3	78.9 / 78	54.2 / 66.8

working_time_correct	AT	BG	FI	FR	IT	PL	SI
NA	0	13.5 / 12.7	0.3 / 0.4	4.3 / 4.7	3.2 / 3.4	0.2 / 0.1	17.9 / 17.1
correct	75.2 / 73.3	62 / 67	67.9 / 76	49 / 54.7	61.6 / 65.5	50.7 / 58.8	55.1 / 67.5
incorrect	24.8 / 26.7	24.6 / 20.3	31.8 / 23.5	46.7 / 40.6	35.2 / 31.1	49.2 / 41.1	26.9 / 15.4

Results 2nd annotation – education, occupation, location

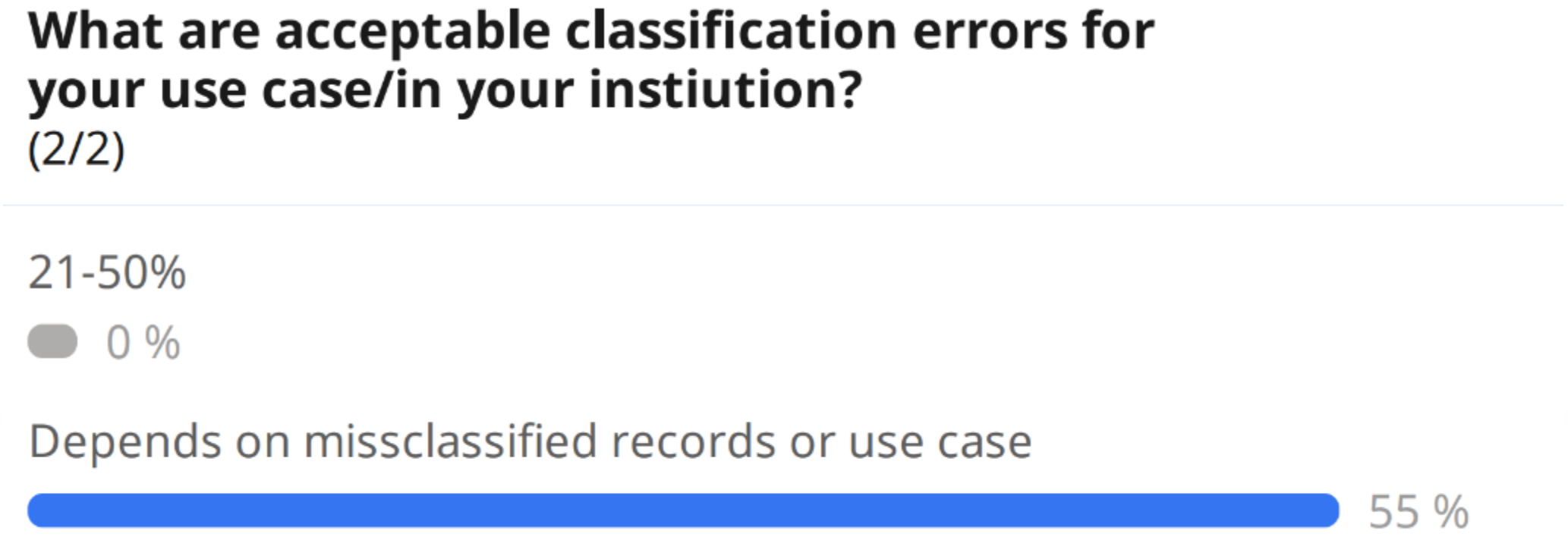
education_correct	AT	BG	FI	FR	IT	PL	SI
NA	0	7.1 / 4.6	0.9 / 0.3	3 / 2.6	3.2 / 2.8	0.3 / 0	15.3 / 9.6
correct	29.5 / 21.5	37 / 51.8	45.2 / 55.5	20.3 / 20.5	12.9 / 13.3	11.8 / 11.7	31.9 / 32.1
incorrect	70.5 / 78.5	55.9 / 43.6	53.9 / 44.2	76.7 / 76.9	83.9 / 83.9	87.9 / 88.2	52.8 / 58.2

occupation1d_correct	AT	BG	FI	FR	IT	PL	SI
NA	0	4 / 3.6	2.4 / 2.7	2.3 / 1.9	2.3 / 2.9	0	12 / 13
correct	49 / 52.7	48.5 / 52.7	62.4 / 63.8	63.3 / 59.7	61.9 / 61	44.7 / 50.1	50.5 / 53.2
incorrect	51 / 47.3	47.5 / 43.8	35.2 / 33.5	34.3 / 38.4	35.8 / 36.2	55.3 / 49.9	37.5 / 33.8

nuts_correct	AT	BG	FI	FR	IT	PL	SI
NA	0	19.2 / 23	0.3 / 0.3	6.3 / 4.2	3.9 / 3	0.3 / 0.3	15.3 / 15.9
correct	75.2 / 73.4	32.3 / 22.1	79.4 / 81	50 / 49.4	58.1 / 58.3	62.6 / 61.7	53.5 / 59.4
incorrect	24.8 / 26.6	48.5 / 54.9	20.3 / 18.7	43.7 / 46.4	38.1 / 38.7	37 / 38	31.2 / 24.7

Slido Question

Slido Question



Slido Question

Do you use other strategies, for assessing the quality of automatic classifications, not mentioned in this webinar?

Various quality measures

implement twice

Cooccurrence analysis

Compare other source

Thank you for joining us today.
If you have any questions, please contacts via email on ESSnet.project@ons.gov.uk

Web Intelligence
Network

Funded by
the European Union