OSCARS
Open Science Clusters' Action for Research & Society

Funded Project

The ONTOLISST project and its preliminary results on using NLP for automated topic assignment

#LoveData25 : AI and Social Sciences Data Webinar, 10 February 2025

Principal Investigator: Judit Gárdos, HUN-REN Centre for Social Sciences
Project partners: Alina Danciu (Sciences Po, France) Mari Kleemola (FSD, Finland),
Miklós Sebők (HUN-REN CSS, Hungary) and their teams

Implemented by

HUN REN Centre for Social Sciences

SciencesPo Centre de Données Socio-Politiques

cnrs

poltextLAB

Tampere University

Finnish Social Science Data Archive

Centre for Social Sciences Research Documentation Centre tkkdk

Funded by the European Union

Project Title: ONTOLISST - Thematic ontologies in social science research data

To improve access to and visibility of social science research

OSCARS Funding:

€ 250,000

Project Start:

December 2024

Project End:

November 2026

Field:

Social Sciences and Humanities

Principal Investigator:

Judit Gárdos, HUN-REN Centre for Social Sciences

Other Researchers involved:

Mari Kleemola, Alina Danciu, Miklós Sebők et al.

Challenge addressed

Diverse metadata ontologies used in social sciences impede data discoverability, accessibility and interoperability.

Step 1	Step 2	Step 3	Step 4	Step 5
Collection of thematic metadata (CVs) from RIs	Create own thesaurus (LiSST)	Create corpus, NLP tests	Research on ontologies	Dissemination, evaluation, demo

IMPACT

ONTOLISST provides tools to simplify topic assignment, thus supporting cross-domain data curation standardization.

Project Tasks and Timeline

Project period: December 2024 - November 2026

Task	Title	Lead	Duration
Task 1	Gathering existing materials: data/metadata/vocabularies	CDSP	M1-M6
Task 2	Creating LiSST: selection of suitable topics for LiSST	HUN-REN CSS KDK	M7-M14
Task 3	Creating the corpus	HUN-REN CSS POL	M15-M16
Task 4	Research on ontologies	HUN-REN CSS KDK	M17-M24
Task 5	NLP experiments: experiment to develop NLP methods to assign LiSST to new data, questions and to variables texts	HUN-REN CSS POL	M17-M22
Task 6	Dissemination: evaluation of results, demo and metadata crosswalk	TAU-FSD	M20-M24
Task 7	Project management	HUN-REN CSS KDK	M1-M24

Project Impacts

OSCARS

Understand curation practices better

- How thematic metadata assignments in social science data archives have worked in the last decades?
- How data curation practices shape social scientific understanding?
- Can lead to more theoretically grounded, well-founded data curation practices in the social science data archival field

Contribute to FAIR social science research data management

- Establish possible ways of assigning thematic metadata more easily, effectively and comparably
- Make data more visible and discoverable, even for visitors without knowledge of the original language of the (meta)data

Build a gold standard corpus with thematic metadata

- Can serve as an important basis and prerequisite to help boost RIs' future abilities to assign thematic metadata more easily using NLP technologies
- Can advance AI-driven metadata curation

Gathering Existing Data/Metadata/Vocabularies

- We will build on existing resources & outputs from various sources
- At the moment, we are still collecting our data:
  - relevant metadata (in DDI format)
  - multilingual controlled vocabularies
- The main thesaurus that we'll use is ELSST – the European Language Social Science Thesaurus
- During the project, we will also explore best ways to share mappings and CV crosswalks

The DDI Alliance is an international membership organization that creates and maintains technical standards for describing research data in the social, demographic, economic, and health sciences.

ELSST Thesaurus

ELSST is a broad-based, multilingual thesaurus for the social sciences. It is owned and published by CESSDA ERIC and its national Service Providers. The thesaurus consists of over 3,400 concepts.

- The following have provided metadata for ONTOLISST (thank you!!)
  - CLOSER UK
  - European Social Survey (Sikt)
  - Generations and Gender Programme
  - GESIS
  - ICPSR
  - Project partners
- Metadata is usually openly available but for this project, we have asked permission to use it.
  - We will not share the metadata outside the project e.g. for other AI tools.
- Do you know of DDI metadata that would include concepts from ELSST or some other (multilingual) thesaurus and that could be used in ONTOLISST? Please let us know!

Project Summary

1. Training an NLP model from on of the provided datasets (CLOSER), to classify the observations into classes/topics
2. Analyzing the XML-DDI files for useful information
3. Predicting with the trained model from the new data from the files

Hungarian Research Network

OSCARS

poltextLAB

HUN-REN CENTRE FOR SOCIAL SCIENCES
BUDAPEST

Political and Legal Text Mining & Artificial Intelligence Laboratory

poltextLAB is a research lab based at the HUN-REN Centre for Social Sciences, Budapest focusing on applications of text mining and artificial intelligence to social science research problems.

LATEST POSTS:

29 JULY, 2024 – MIKLÓS SEBŐK AMONG NRDI NREP EXCELLENCE PROGRAMME WINNERS

20 JUNE, 2024 – REBEKA KISS'S PRESENTATION ON THE QUALITY OF LEGISLATION

18 JUNE, 2024 – SEAN THERIAULT'S LECTURE ON THE POLICY AGENDAS OF THE HOLY SEE AS A GLOBAL POWER

poltextLAB

SHIFT

ARTIFICIAL INTELLIGENCE National Laboratory

1. Training NLP model to text classification

- **Task:**

  building a representative example of automated text classification with NLP models, with an already existing codebook

- **Source material:**

  UK survey titles and questions coded with the CLOSER codebook, ~20k unique observations, in English

- **16 Main Topics:**
  a. Topic code:
     original CLOSER codes
  b. Topic name
  c. Code:
     codes we used in the
     classification task (labels)

Topic code	Topic name	Code
101	Demographics	0
102	Housing and local environment (Housing and environment)	1
103	Physical health	2
104	Mental health and mental processes	3
105	Health care	4
106	Health behaviour (Health and lifestyle)	5
107	Family and social networks	6
108	Education	7
109	Employment and income (Employment and pensions)	8
110	Expectations, attitudes and beliefs (Attitudes and beliefs)	9
111	Child development	10
112	Life events	11
113	Omics	12
114	Pregnancy	13
115	Administration	14
116	COVID19	15

- **119 Subtopics (examples):**
  a. Education subtopics →
  b. Health care subtopics ↓

Topic code	Topic name	Code
105	Health care	4
10501	Health services utilisation	55
10502	Hospital admissions	56
10503	Immunisations	57
10504	Medications	58
10505	Complementary therapies	59
10506	Health insurance	60

Topic code	Topic name	Code
108	Education	7
10801	Qualifications	81
10803	Further education | Higher education	82
10804	Training	83
10805	Basic skills	84
10806	Adult education	85
10807	Learning difficulties	86
10808	Pre-school	87
10809	Cognitive function	88
10810	Cognitive skills	89
10811	Non cognitive skills	90
10812	School engagement	91
10813	Education aspirations	92
10814	Lifelong learning	93
10815	Primary schooling	94
10816	Secondary schooling	95

id	text	label1
qi_H1_c	This questionnaire was completed by: (tick all that apply) other (please describe)	0
qi_H1_b	This questionnaire was completed by: (tick all that apply) father	0
qi_B3	What is today's date?	0

Code 0 = Demographics

qi_E36_b	How many drinks containing alcohol do you have on a typical day when you are drinking?	5
qi_E36_h	How often during the last year have you been unable to remember what happened the night before because you had been drinking?	5
qi_E36_d	How often during the last year have you found that you were not able to stop drinking once you had started?	5
qi_D19_i	What about your mother's or your father's views on your drinking? Mother	5

Code 5 = Health behavior

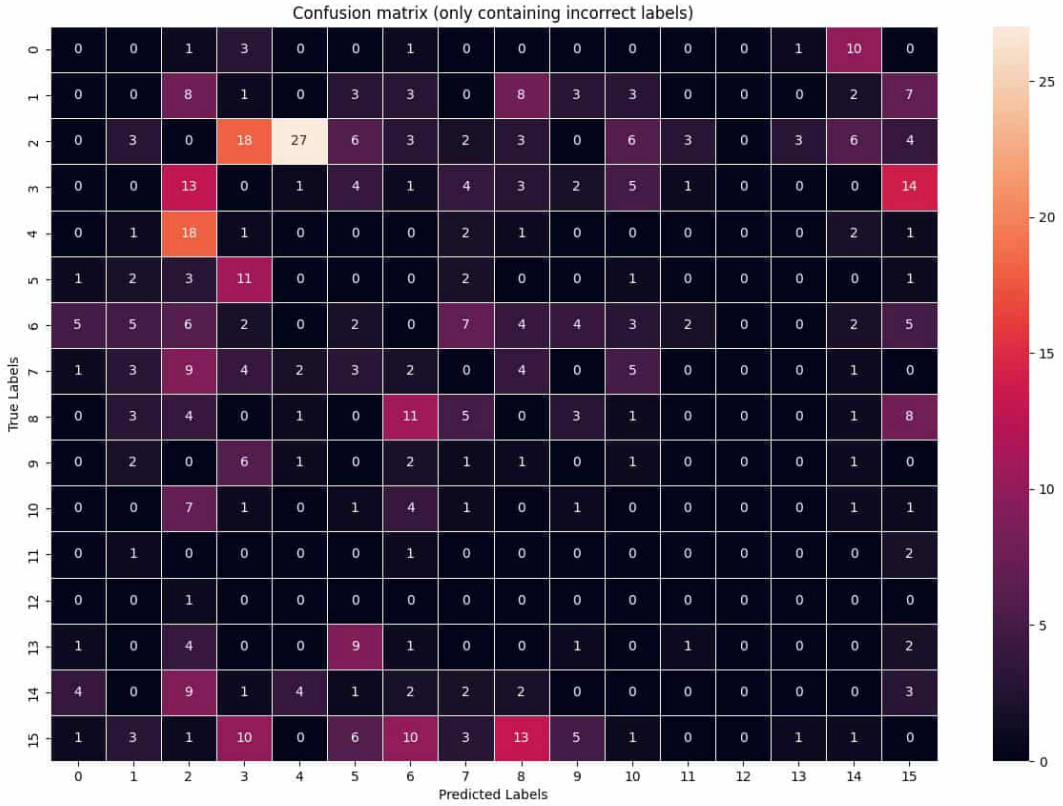
- **Model training:**
  a. Training an XLM-Roberta multilingual large language model on the dataset, which contains question texts and the related labels from CLOSER
  b. Training data: 16 901 observations, test data: 4 216 observations
- **Model versions:**
  a. **Version 1**: training with the main topics
  b. **Version 2**: training with the subtopics

poltextLAB Demo Project for ONTOLISST - Results

OSCARS

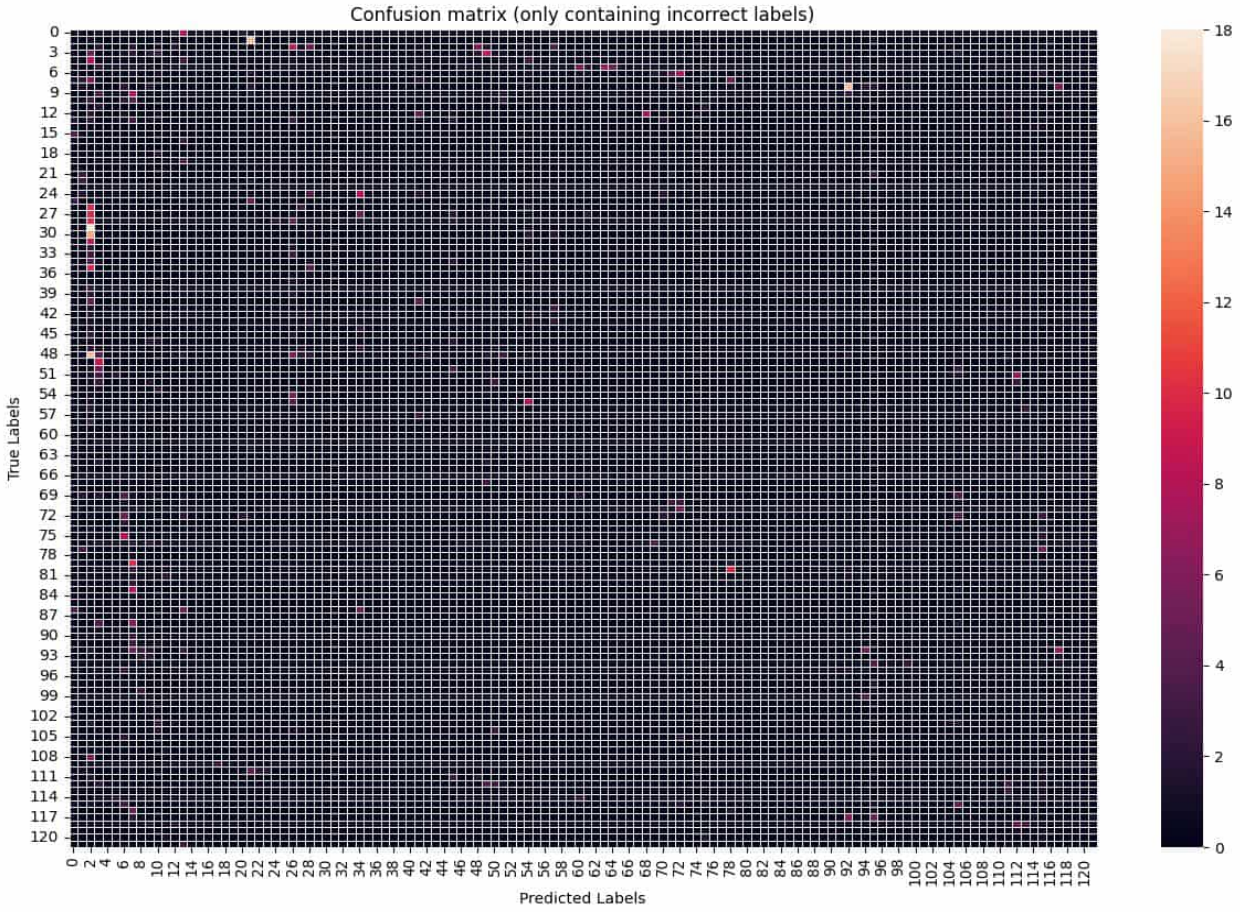
Model V1 for 16 major topics:

- Model V1 for 16 major topics
- F1 score: 0.88
- Human-level score - infinitely replicable within seconds
- Most bad labels for health related topics



class	precision	recall	f1-score	support
0	0.88	0.86	0.87	111.00
1	0.88	0.82	0.85	207.00
2	0.89	0.89	0.89	789.00
3	0.89	0.90	0.90	504.00
4	0.67	0.73	0.70	98.00
5	0.89	0.93	0.91	297.00
6	0.90	0.89	0.89	416.00
7	0.90	0.89	0.89	301.00
8	0.91	0.91	0.91	426.00
9	0.77	0.81	0.79	79.00
10	0.90	0.93	0.91	244.00
11	0.86	0.92	0.89	48.00
12	0.00	0.00	0.00	1.00
13	0.89	0.69	0.78	61.00
14	0.76	0.75	0.76	113.00
15	0.91	0.89	0.90	521.00
accuracy	0.88	0.88	0.88	0.88
macro avg	0.81	0.80	0.80	4216.00
weighted avg	0.88	0.88	0.88	4216.00

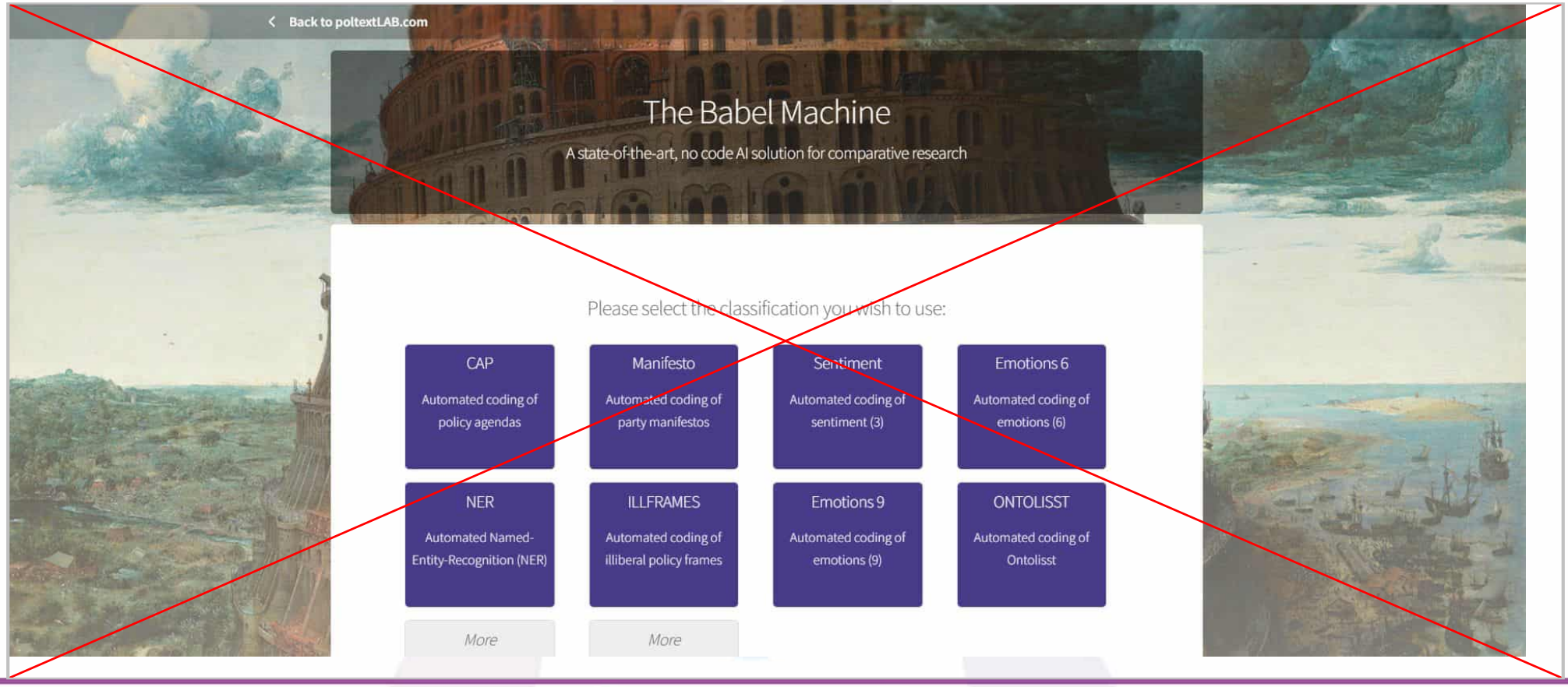
Model V2 for 119 subtopics:



class	precision	recall	f1-score	support
0	0.69	0.70	0.70	54.0
1	0.26	0.23	0.24	26.0
2	0.38	0.72	0.49	157.0
3	0.80	0.87	0.83	254.0
4	0.00	0.00	0.00	19.0
...	...	...	...	...
133	0.00	0.00	0.00	0.0
134	0.81	0.78	0.79	27.0
micro avg	0.69	0.69	0.69	4215.0
macro avg	0.40	0.39	0.38	4215.0
weighted avg	0.64	0.69	0.65	4215.0

Model Version 1 is now available in the BABEL Machine:

https://ontolisstbabel.poltextlab.com/



This is what the automated email looks like:

Dear Jane Doe!

The Babel Machine has processed your request successfully. Click https://storage.cloud.google.com/cap-babel-eu/results/ess_sample_lULNJP38Q5_with_pred.csv to download the coded dataset [Sample] in CSV format.
You will find the predicted major topic codes in the new column: "major_topic_pred". Please use tools such as pandas to open the file as spreadsheet editors might have issues with correctly displaying the file contents. Please note that your dataset may have less rows compared to your initial submission if we have detected empty rows in your dataset, or if you had any cells in the text column with exactly "NA" as the value. We treat these as actual NA values and therefore we will not process them.

The language models that The Babel uses were fine-tuned on training data containing only the Main-Categories of the coding scheme (the three-digit variables), excluding the Sub-Categories (the four-digit variables and the four digit variables with underscore). We use the label 0 to indicate that the given text contains no meaningful information.

You can download a file containing the three highest probability category prediction by The Babel model and the corresponding probability (softmax) score assigned to each label. To download it please use this link: https://storage.cloud.google.com/cap-babel-eu/softmax/ess_sample_lULNJP38Q5_softmax.csv. If your dataset had text where the value was exactly "NA", then the corresponding row(s) in the softmax sheet will be empty, as NA values do not have assigned prediction probabilities.

Cost of running the process (excluding support and development costs): $20.59

Please note that this e-mail was automatically generated by The Babel Machine.
If you have any questions do not hesitate to contact us at poltextlab@poltextlab.com.

This is what the automated email contains as an output csv (sample):

text	major_topic_pred	major_topic_pred_name
TV watching, total time on average weekday	5	Health behaviour (Health and lifestyle)
Most of the time people helpful or mostly looking out for themselves	3	Mental health and mental processes
You think employer considered job to be temporary or permanent	8	Employment and income (Employment and pensions)
Eighteenth person in household: Relationship to respondent	6	Family and social networks
Year of birth	0	Demographics
Allowed to choose/change pace of work	9	Expectation, attitudes and beliefs (Attitudes and beliefs)
Year of birth of twenty-second person in household	0	Demographics
Age of respondent, calculated	0	Demographics
TV watching, total time on average weekday	5	Health behaviour (Health and lifestyle)
TV watching, news/politics/current affairs on average weekday	5	Health behaviour (Health and lifestyle)
Main reason for leaving last employer	8	Employment and income (Employment and pensions)
Radio listening, news/politics/current affairs on average weekday	5	Health behaviour (Health and lifestyle)
Newspaper reading, politics/current affairs on average weekday	5	Health behaviour (Health and lifestyle)
Most people can be trusted or you can't be too careful	3	Mental health and mental processes
Most people try to take advantage of you, or try to be fair	9	Expectation, attitudes and beliefs (Attitudes and beliefs)
Most of the time people helpful or mostly looking out for themselves	3	Mental health and mental processes
Paid employment or apprenticeship at least 3 months 20 hours weekly	8	Employment and income (Employment and pensions)
Year first started in paid employment or apprenticeship	8	Employment and income (Employment and pensions)
Proportion of household income respondent provides	8	Employment and income (Employment and pensions)
Year first left parents for living separately for 2 months or more	10	Child development

2. Creating compatibility between DDI Files and the Babel AI models

- **Now in DDI files:**
  - Keywords, concepts assigned to surveys
  - Observations ("Variables") ordered into classes ("Variable Groups")
  - No standard codebook for this ➔ *LISST*
- **Aim:**

  Integrating the DDI standard files into the model

- **First test:**

  How well does the question-trained model work on "Variables"?

- **Example:**

Question text:		DDI Variable:
“Can you tell me, what was [child name]’s age last birthday?”	↔	“Child age exact”

3. Predicting with the question-trained model

- **Task:**
  - Classifying the observations (Variables) with the question-trained model (16 classes) and compare the predicted labels with the original classes (Variable Groups) from the files
  - SIKT - ESS: 20 classes, GESIS: 21 classes
- **From the results:**
  - We can see that the model stays consistent and assigns similar topics to the observations (next slides)

ESS Topics vs. Predicted CLOSER Labels

OSCARS

Most dominant predicted labels for ESS topics examples:

0: 'Media use and trust' - Health and Lifestyle,COVID,Physical Health,Mental Health

1: 'Justice' - Attitudes and Beliefs,

2: 'Personal and household characteristics' - Demographics,Family and Social Networks,Pregnancy,

3: 'Family, work and wellbeing' - Employment,Family and Social Networks, Attitudes and Beliefs,Education,

4: 'Human values' - Attitudes and Beliefs,

5: 'Immigration' - Attitudes and Beliefs,Demographics,

6: 'Climate change' - Attitudes and Beliefs,

7: 'Welfare attitudes' - Attitudes and Beliefs,Employment,

8: 'Democracy' - Attitudes and Beliefs

9: 'Personal and social wellbeing' - Mental Health,

10: 'Timing of life' - Life Events,Family and Social Networks,Attitudes and Beliefs,

11: 'Justice and fairness' - Employment,Attitudes and Beliefs,

Most dominant labels for GESIS topics examples:

0: 'General interest services - use and experience' - Administration, Housing, Physical health

1: 'General interest services - Consumer protection' - COVID19

2: 'Demographics 1' - Employment

3: 'Demographics All' - Housing,Family and social networks,Employment

4: 'Employment and social policy 1' - Employment,COVID19

5: 'Employment opportunities and training' - Employment

6: 'Employment and social policy 2' - Employment

7: 'National policy items' - Administration

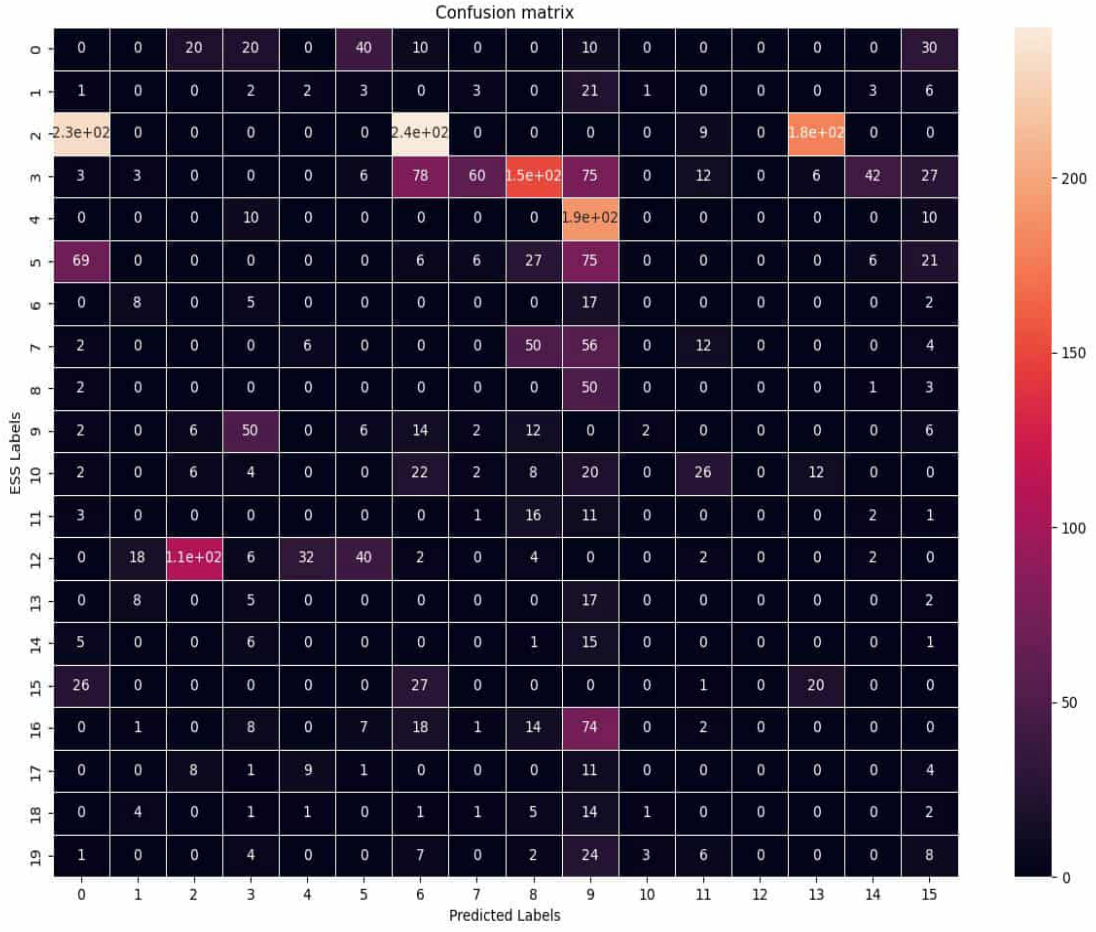
8: 'Energy - technologies' - Housing

9: 'Energy - policy' - Housing,Attitudes and beliefs

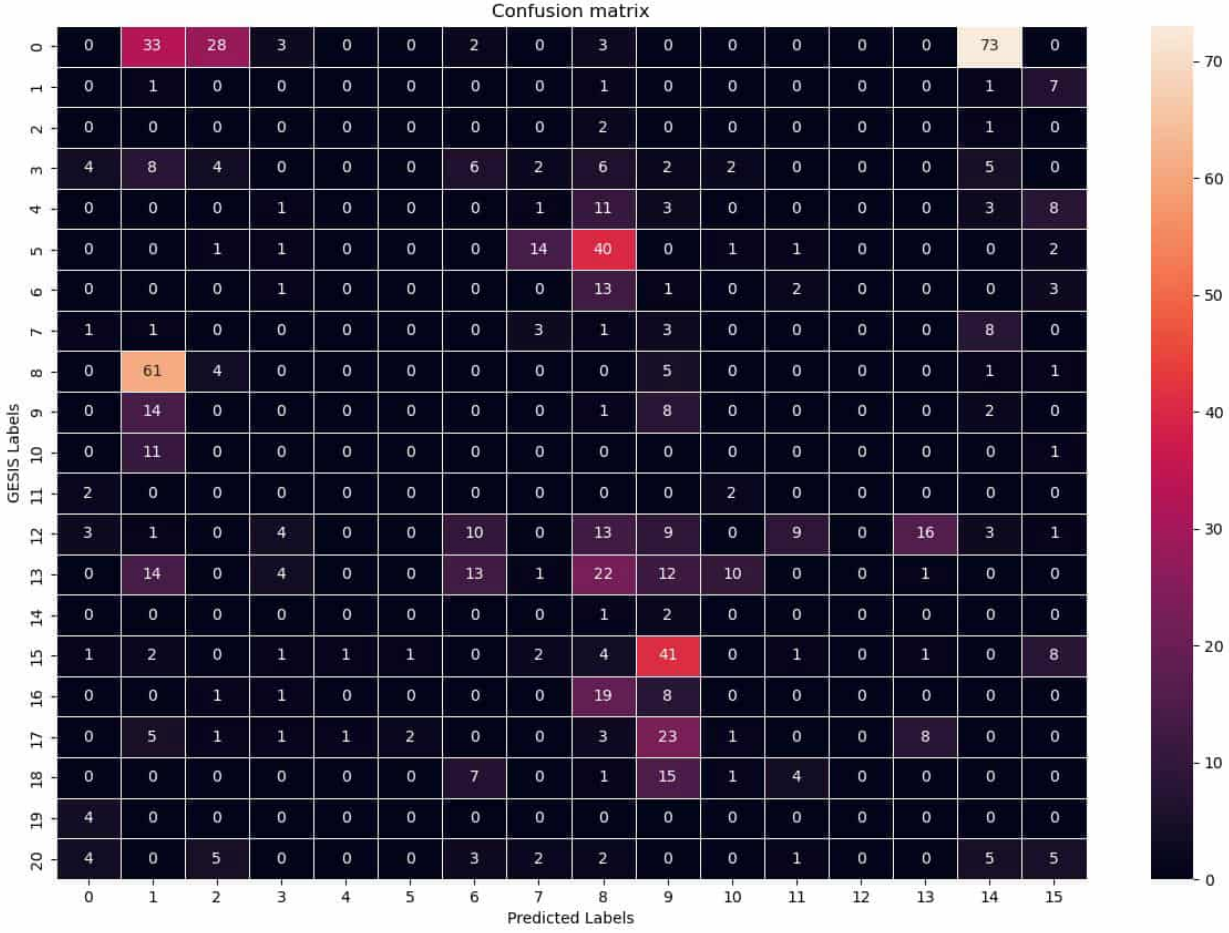
10: 'Energy - consumption and saving' - Housing

poltextLAB Demo Project for ONTOLISST - Version 1 Results

SIKT - ESS Variable groups



GESIS Variable groups



Conclusions

1. Question-trained NLP models can accurately recognise concepts in survey questions from a given codebook ("variable groups", "labels")
2. The question-trained model can also classify "DDI standard Variables"

Benefits for the project

1. Before the final *LISST* codebook ("ontology") is available: empirically tested ideas for the incorporation of already used "variable groups", "concepts" as labels
2. After the final *LISST* codebook: the final trained AI model will be able to automatically code question texts,"variables" fast and with good accuracy

More information:

https://oscars-project.eu/projects/ontolisst-thematic-ontologies-social-science-research-data


OSCARS
Thank you