DiRAC Federation Project
Final Digital Asset - Report on Results

Work Package/Activity: D-FED 3.1.7 - Understanding the multiple dimensions of prediction of concepts in social and biomedical science questionnaires

Authors:
Suparna De (Work Package/Activity Lead)
Jon Johnson (Work Package/Activity Co-Lead)
Jenny Li (Developer)
Harry Moss (Developer)
Sanaz Jabbari (Analyst)

Introduction

The CLOSER Discovery (Discovery, 2021) initiative provides metadata to support the emerging needs of data and variable-level discovery for the UK's longitudinal studies. It aims to develop an ecosystem to support the long-term viability of producing high granularity metadata that captures the data and survey data collection lifecycle.

The content of the surveys in these longitudinal studies are in the form of PDF questionnaires with different semantically distinct elements (question text, conditions, instructions etc.) that are captured within the DDI-Lifecycle standard schema. The application of common ontologies to a data collection is a significant investment to aid discovery and uptake of data for secondary data analysis. As data scales and ontologies develop, this is a major burden on providing ease of entry to using data investments. Existing efforts within CLOSER Discovery have involved manual and semi-automated tagging of question items to the CLOSER ontology, which have been utilised by the RCNIC-funded project 'Automated classification of social science questions' as training data to explore machine learning (ML) algorithms for the automated tagging of question items to existing thesauri.

This work package on 'Understanding the multiple dimensions of prediction of concepts in social and biomedical science questionnaires' extends the scope of the research tackled in the RCNIC project to:

1. dive deeper into questions related to the size and quality of the training data and how this affects the performance of the designed ML models,
2. assess the performance of the trained ML models for automated tagging of question texts with the top-level concept topics (14 in number) from existing thesauri such as European Language Social Science Thesaurus (ELSST) in 'inference mode', i.e. with new unseen questionnaires (that were not part of the training and validation set)
3. investigate new ML models (such as hierarchical approaches) for tagging question texts (and response domains) with the 120 second-level topics from ELSST.

Machine Learning Pipeline

The project has developed a ML pipeline, which enables running a large number of combinations of ML models and optimisations, different combinations of data through to the output measurement metrics.

Git is used for version control of the underlying code used to pre-process input data, generate features for training from the data and for model training and evaluation. Additionally, text outputs of experiments and basic plots are versioned with Git. In this structure, each broad model family occupies a branch, with individual experiments represented by a directory containing output files following model training and evaluation.

The ML pipeline relies upon DVC (Kuprieiev, 2021) to perform both dataset versioning and MLFlow for experiment tracking. In this context, an experiment refers to the model training process: from the choice of input parameters to the performance of the trained model on validation data according to several metrics of interest, such as accuracy, precision, recall, f1-score and the area under the receiver operating characteristic curve, referred to here as AUC score and ROC curve. Datasets versioned by DVC are referenced in the version-controlled codebase, managed by Git, and transferred via Secure Shell (SSH) to remote storage on the University College London (UCL) Research Data Storage Service (UCL, 2021). Full details of the ML pipeline design and working are documented in (De et al., 2022).

Work Package Deliverables

1. Concept prediction - understanding different predictions rate by category

Concept prediction for the question texts for the 14 top-level topics was investigated in the RCNIC project. The problem was cast as a classification problem using supervised learning. Four broad model architectures were considered and performance compared:

- Multinomial naïve Bayes (MNB) model - selected to determine a performance baseline against which other models with a greater number of parameters and level of complexity are compared.
- Long short term memory (LSTM) model - implementation of a deep neural network model architecture, particularly suited to processing sequential data by extending the architecture of recurrent neural networks (RNNs).
- Universal Language Model Fine-tuning for Text Classification (ULMFit; Howard and Ruder, 2018) - LSTM enhancement with a language model pretrained on Wikitext-103 (Merity et al., 2017b), a general English language corpus containing over 100 million words extracted from quality-assured Wikipedia articles. ULMFit utilises a transfer learning approach, with the intention that the weights of this language model be *fine-tuned* on the corpus of training data.
- BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018) - class of models based on the *transformer,* using *self-attention* to adaptively weight sections of input data by significance.

After assessing performance using the baseline Multinomial naive Bayes classifier, the performance of three neural network architectures, described above, was assessed in comparison. The performance of all three neural network architectures was assessed through a hyperparameter tuning exercise in which a grid of hyperparameters was generated. In this case, hyperparameters considered were learning rate, batch size, metadata addition and optimiser. In turn, experiments were generated by sampling from the grid of hyperparameters, enabling a set of optimal hyperparameters to be determined for each model.

In addition to hyperparameter optimisation, several experiments were conducted to investigate the addition of questionnaire metadata into the features used for model training. An initial approach looked at concatenating the question literal string (an individual sample in the original approach) with the question response string. After observing improved performance with question-reponse concatenation, questionnaire instrument name and questionnaire section heading were considered as additional candidates for inclusion in training features. Through concatenation with the question literal string, both metadata additions yielded increased model performance, with the section heading metadata maximally improving performance.

For all models, the performance metric of interest was chosen to be the so-called F1-score, the harmonic mean of precision and recall, where

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

as shown in equation 1. For aggregate measures, a simple mean F1 score calculated from the F1 score of all classes is provided by the “Macro average” entry in tables 1-4. A weighted average, taking into account the number of samples in each category, is also given and is the preferred aggregate measure for model performance measure and model comparison.

Figure 1 shows the performance via weighted average f1-score of the various models with question-response concatenation and the addition of section heading metadata for the individual top-level topics. The topic of ‘life events’ is clearly distinguishable from the rest of the topics as having tightly grouped f1-scores at high values in all models. This is followed by ‘health behaviour’, ‘Employment and income’, ‘Mental health and mental processes’, ‘Physical health’, which also get good performance, particularly from the BERT and ULMFit models. Tables 1 to 4 show the f1-scores of the top-level concept classification under the “14-class f1-score” heading.

In aggregate and when considered on a class-by-class basis, ULMFit (with the combination of hyperparameters and metadata chosen here) is shown to have the greatest performance by f1-score. This is closely followed by BERT base uncased, although this has a greater variability in performance across classes. Of the three neural network architectures, our “Simple LSTM” model consisting of an LSTM cell, a dropout layer and a linear layer is considered to be the simplest implementation of all three, although it outperforms or comes close to the performance of BERT in several classes. All of the neural network-based models outperform the baseline Multinomial naive Bayes, justifying to some degree the greater computational effort required to train the models.

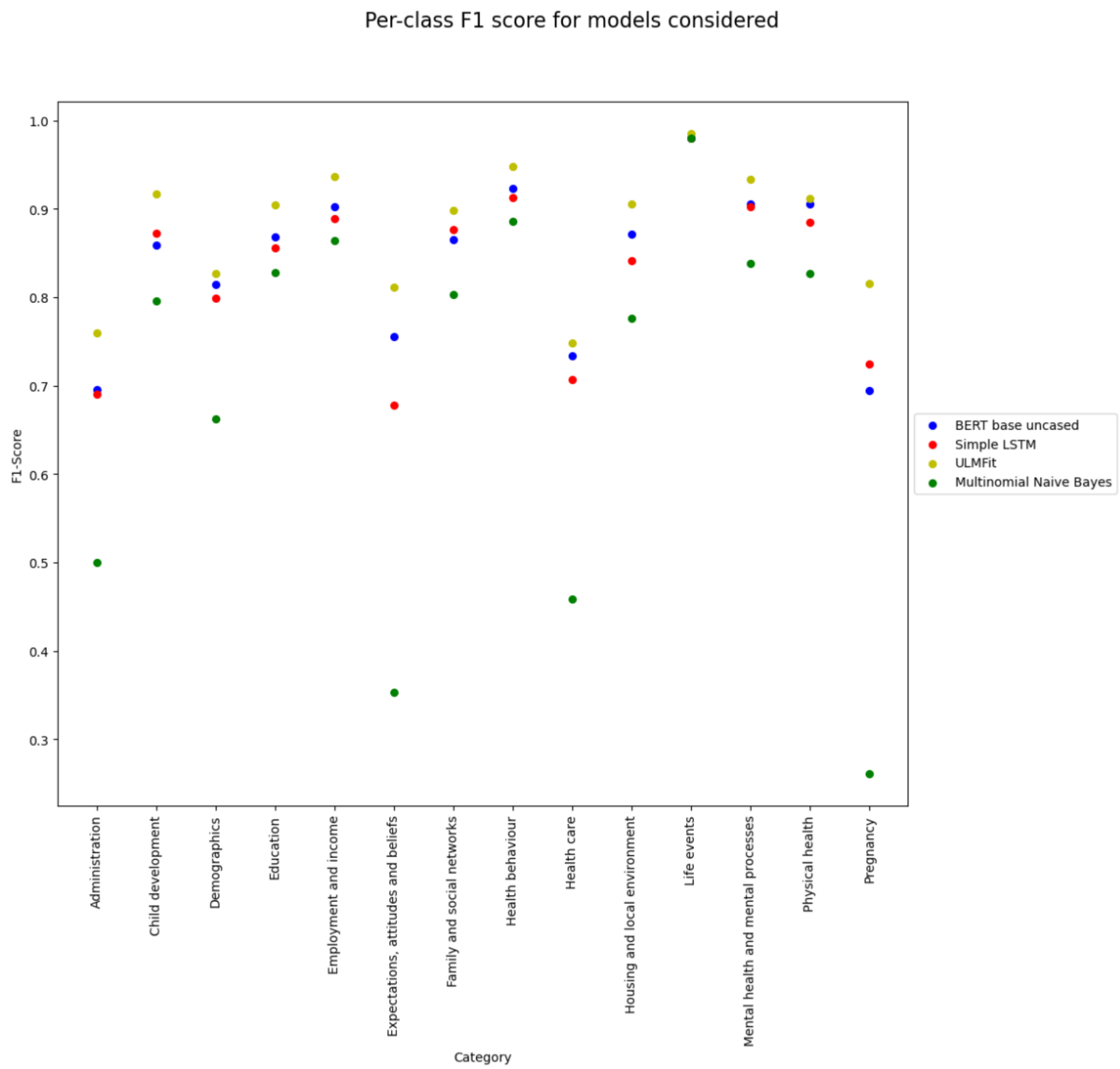


Fig.1. Per-class f1-score for each considered model. All models are trained with question-response concatenation and the inclusion of section heading metadata, as these were observed to produce the most performant models in each case.

All model architectures were then trained to classify questions as one of up to 100 second-level topics from the ELSST ontology. The second-level topic classification experiment required a separate pre-processing stage to generate a training dataset with extracted second level topic labels from the 5-digit numerical representation of the question topic, where the first 3 represent the top-level concept and the last two the second-level concept. In the absence of a 5-digit numerical representation, the top-level 3-digit representation was used. Classes containing 5 or fewer examples were discarded.

Results for the second-level concept are provided, alongside the counterpart top-level result, for each model in tables 1-4. Figures 2 to 5 visually represent the per-class f1-score for both the top-level 14-class problem and the second-level 100-class problem. In all cases, the model variant including section heading metadata in training features was found to yield the highest aggregate F1 score.

Table 1: Second level per-class and aggregate f1-scores (in bold) for the BERT_base_uncased model trained with question-response concatenation and the addition of section heading metadata. Note that the 100-class categorisation model includes the "COVID-19" and "Omics" classes, which are not shown here.

Category	100-class CODE	100-class f1-score	support	14-class CODE	14-class - f1-score	dataset size
Place of birth	10101	0.889	15	101	0.814	203
Gender	10102	0.710	18	101	0.814	203
Ethnic group	10103	0.977	21	101	0.814	203
Language(s) spoken	10104	0.947	10	101	0.814	203
Location	10106	0.757	21	101	0.814	203
Age	10107	0.313	20	101	0.814	203
Housing	10201	0.785	137	102	0.871	355
Neighbourhood	10202	0.881	43	102	0.871	355
Travel and transport	10203	0.879	64	102	0.871	355
Environmental exposure	10204	0.840	77	102	0.871	355
Cardiovascular system	10301	0.761	83	103	0.906	1372
Musculoskeletal system	10302	0.714	81	103	0.906	1372
Nervous system	10304	0.624	46	103	0.906	1372
Digestive system	10305	0.724	60	103	0.906	1372
Urogenital system	10306	0.796	52	103	0.906	1372
Endocrine system	10307	0.800	16	103	0.906	1372
Hemic and immune systems	10308	0.625	10	103	0.906	1372
Hearing, vision, speech	10309	0.852	137	103	0.906	1372
Oral/dental health	10310	0.822	52	103	0.906	1372
Congenital malformations	10312	0.667	10	103	0.906	1372
Cancer	10313	0.000	4	103	0.906	1372
Mortality	10314	0.519	17	103	0.906	1372
Women's health	10316	0.787	44	103	0.906	1372
Accidents and injuries	10317	0.920	83	103	0.906	1372
Allergies	10318	0.853	70	103	0.906	1372
Infections	10319	0.882	34	103	0.906	1372
Anthropometry	10320	0.879	100	103	0.906	1372
Physical characteristics	10321	0.831	41	103	0.906	1372
Physical functioning	10322	0.686	54	103	0.906	1372
General health	10323	0.462	138	103	0.906	1372
Mental disorders	10401	0.623	32	104	0.905	942
Personality | Temperament	10402	0.803	238	104	0.905	942
Wellbeing	10403	0.783	51	104	0.905	942
Emotions	10404	0.612	47	104	0.905	942
Cognitive function	10405	0.685	41	104	0.905	942

Health services utilisation	10501	0.491	64	105	0.734	251
Hospital admissions	10502	0.619	72	105	0.734	251
Immunisations	10503	0.526	10	105	0.734	251
Medications	10504	0.743	38	105	0.734	251
Complementary therapies	10505	0.952	11	105	0.734	251
Diet and nutrition	10601	0.932	256	106	0.923	542
Physical activity	10602	0.741	56	106	0.923	542
Alcohol consumption	10605	0.959	84	106	0.923	542
Substance abuse	10606	0.981	52	106	0.923	542
Criminal behaviour	10608	0.750	5	106	0.923	542
Home life	10701	0.737	78	107	0.865	823
Household composition	10702	0.708	83	107	0.865	823
Marital status	10703	0.769	67	107	0.865	823
Family members and relations	10704	0.682	152	107	0.865	823
Friends	10705	0.667	29	107	0.865	823
Childcare	10706	0.717	27	107	0.865	823
Child welfare	10707	0.000	9	107	0.865	823
Social support	10708	0.775	104	107	0.865	823
Leisure activities	10709	0.790	93	107	0.865	823
Technology	10711	0.828	17	107	0.865	823
Qualifications	10801	0.904	95	108	0.869	617
Further education | Higher education	10803	0.605	38	108	0.869	617
Training	10804	0.706	25	108	0.869	617
Basic skills	10805	0.797	57	108	0.869	617
Adult education	10806	0.000	7	108	0.869	617
Learning difficulties	10807	0.577	28	108	0.869	617
Pre-school	10808	0.667	6	108	0.869	617
Cognitive skills	10810	0.759	17	108	0.869	617
Non cognitive skills	10811	0.000	5	108	0.869	617
Education aspirations	10813	0.522	15	108	0.869	617
Primary schooling	10815	0.780	21	108	0.869	617
Occupation | Employment	10901	0.891	385	109	0.903	711
Social classification	10902	0.400	15	109	0.903	711
Income	10903	0.843	62	109	0.903	711
Finances	10904	0.793	57	109	0.903	711
Assets	10905	0.333	5	109	0.903	711
Consumption | Expenditure	10906	0.711	23	109	0.903	711
Pensions	10907	0.000	2	109	0.903	711
Benefits | Welfare	10908	0.833	13	109	0.903	711
Social attitudes	11001	0.750	5	110	0.755	126
Politics	11002	0.857	19	110	0.755	126
Infant feeding	11101	0.806	34	111	0.859	585

Language and vocabulary	11102	0.911	83	111	0.859	585
Parenting	11103	0.811	186	111	0.859	585
Developmental milestones	11104	0.864	63	111	0.859	585
Childbirth	11401	0.658	34	114	0.694	119
Macro average		**0.705**	7038		**0.841**	7034
Weighted average		**0.781**	7038		**0.875**	7034

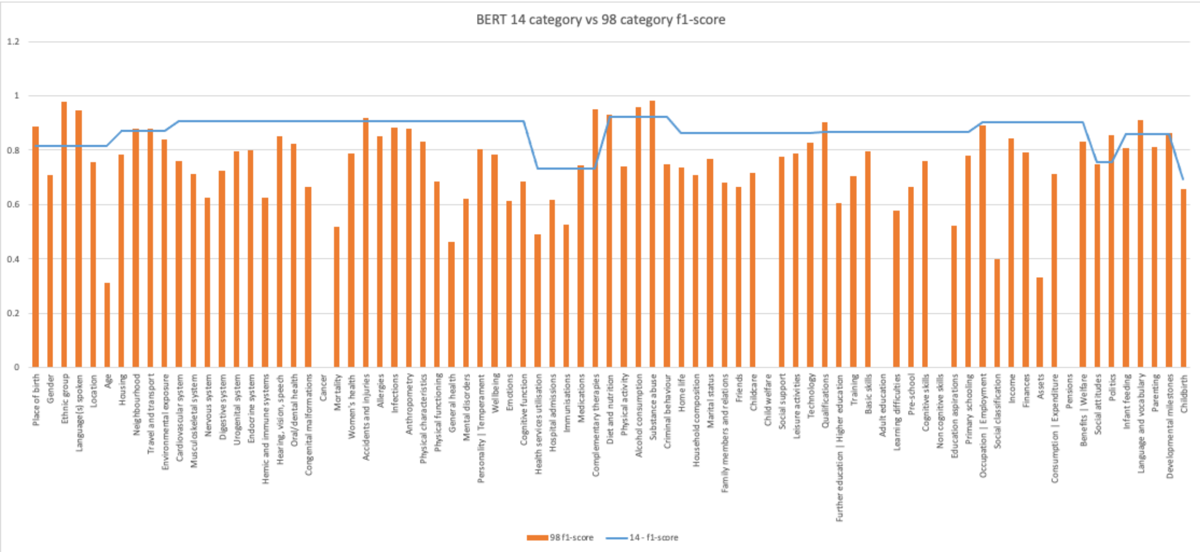


Fig.2. Per-class F1-scores for BERT_base_uncased in both 14-class and 98-class classification.

Table 2: Second level per-class and aggregate (in bold) f1-scores for the ULMFit model trained with question-response concatenation and the addition of section heading metadata. Note that the 100-class categorisation model includes the "COVID-19" and "Omics" classes, which are not shown here.

Category	98 CODE	98- f1-score	# Items (98-class)	14 CODE	14 - f1-score	# Items (98-class)
Place of birth	10101	0.769	15	101	0.827	203
Gender	10102	0.105	18	101	0.827	203
Ethnic group	10103	0.744	21	101	0.827	203
Language(s) spoken	10104	0.000	10	101	0.827	203
Location	10106	0.087	21	101	0.827	203
Age	10107	0.000	20	101	0.827	203
Housing	10201	0.713	137	102	0.905	355
Neighbourhood	10202	0.627	43	102	0.905	355
Travel and transport	10203	0.794	64	102	0.905	355
Environmental exposure	10204	0.696	77	102	0.905	355
Cardiovascular system	10301	0.497	83	103	0.912	1372
Musculoskeletal system	10302	0.358	81	103	0.912	1372

Nervous system	10304	0.328	46	103	0.912	1372
Digestive system	10305	0.667	60	103	0.912	1372
Urogenital system	10306	0.545	52	103	0.912	1372
Endocrine system	10307	0.000	16	103	0.912	1372
Hemic and immune systems	10308	0.000	10	103	0.912	1372
Hearing, vision, speech	10309	0.598	137	103	0.912	1372
Oral/dental health	10310	0.634	52	103	0.912	1372
Congenital malformations	10312	0.000	10	103	0.912	1372
Cancer	10313	0.000	4	103	0.912	1372
Mortality	10314	0.000	17	103	0.912	1372
Women's health	10316	0.564	44	103	0.912	1372
Accidents and injuries	10317	0.888	83	103	0.912	1372
Allergies	10318	0.761	70	103	0.912	1372
Infections	10319	0.724	34	103	0.912	1372
Anthropometry	10320	0.783	100	103	0.912	1372
Physical characteristics	10321	0.737	41	103	0.912	1372
Physical functioning	10322	0.450	54	103	0.912	1372
General health	10323	0.305	138	103	0.912	1372
Mental disorders	10401	0.516	32	104	0.934	942
Personality | Temperament	10402	0.699	238	104	0.934	942
Wellbeing	10403	0.355	51	104	0.934	942
Emotions	10404	0.259	47	104	0.934	942
Cognitive function	10405	0.483	41	104	0.934	942
Health services utilisation	10501	0.242	64	105	0.748	251
Hospital admissions	10502	0.663	72	105	0.748	251
Immunisations	10503	0.000	10	105	0.748	251
Medications	10504	0.559	38	105	0.748	251
Complementary therapies	10505	0.778	11	105	0.748	251
Diet and nutrition	10601	0.840	256	106	0.948	542
Physical activity	10602	0.667	56	106	0.948	542
Alcohol consumption	10605	0.800	84	106	0.948	542
Substance abuse	10606	0.868	52	106	0.948	542
Criminal behaviour	10608	0.000	5	106	0.948	542
Home life	10701	0.614	78	107	0.899	823
Household composition	10702	0.685	83	107	0.899	823
Marital status	10703	0.587	67	107	0.899	823
Family members and relations	10704	0.575	152	107	0.899	823

Friends	10705	0.065	29	107	0.899	823
Childcare	10706	0.138	27	107	0.899	823
Child welfare	10707	0.000	9	107	0.899	823
Social support	10708	0.650	104	107	0.899	823
Leisure activities	10709	0.638	93	107	0.899	823
Technology	10711	0.000	17	107	0.899	823
Qualifications	10801	0.887	95	108	0.904	617
Further education | Higher education	10803	0.190	38	108	0.904	617
Training	10804	0.000	25	108	0.904	617
Basic skills	10805	0.686	57	108	0.904	617
Adult education	10806	0.000	7	108	0.904	617
Learning difficulties	10807	0.067	28	108	0.904	617
Pre-school	10808	0.000	6	108	0.904	617
Cognitive skills	10810	0.000	17	108	0.904	617
Non cognitive skills	10811	0.000	5	108	0.904	617
Education aspirations	10813	0.000	15	108	0.904	617
Primary schooling	10815	0.000	21	108	0.904	617
Occupation | Employment	10901	0.761	385	109	0.937	711
Social classification	10902	0.000	15	109	0.937	711
Income	10903	0.694	62	109	0.937	711
Finances	10904	0.672	57	109	0.937	711
Assets	10905	0.000	5	109	0.937	711
Consumption | Expenditure	10906	0.545	23	109	0.937	711
Pensions	10907	0.000	2	109	0.937	711
Benefits | Welfare	10908	0.526	13	109	0.937	711
Social attitudes	11001	0.000	5	110	0.811	126
Politics	11002	0.773	19	110	0.811	126
Infant feeding	11101	0.227	34	111	0.917	585
Language and vocabulary	11102	0.827	83	111	0.917	585
Parenting	11103	0.678	186	111	0.917	585
Developmental milestones	11104	0.812	63	111	0.917	585
Childbirth	11401	0.592	34	114	0.815	119
Macro average		**0.440**	7038		**0.879**	7034
Weighted average		**0.627**	7038		**0.904**	7034

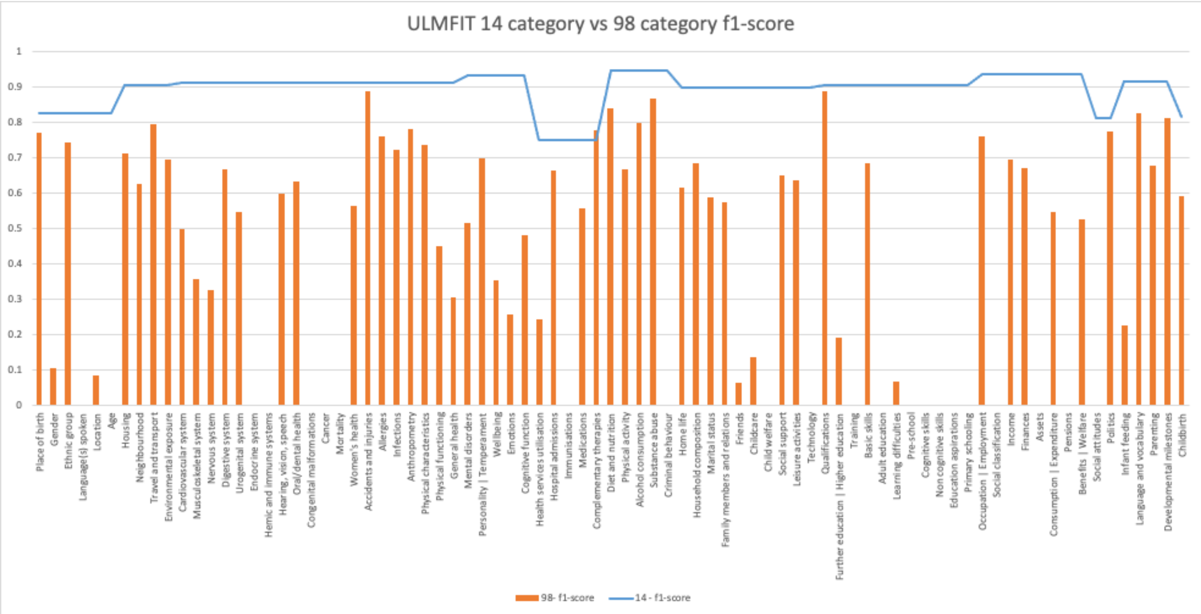


Fig.3. Per-class F1-scores for ULMFit in both 14-class and 98-class classification.

Table 3: Second level per-class and aggregate (in bold) f1-scores for the Simple LSTM model trained with question-response concatenation and the addition of section heading metadata. Note that the 100-class categorisation model includes the "COVID-19" and "Omics" classes, which are not shown here.

Category	98 CODE	98 f1-score	support	14 CODE	14 - f1-score	dataset size
Place of birth	10101	0.786	13	101	0.798	194
Gender	10102	0.692	15	101	0.798	194
Ethnic group	10103	0.913	22	101	0.798	194
Language(s) spoken	10104	0.706	8	101	0.798	194
Location	10106	0.686	20	101	0.798	194
Age	10107	0.686	19	101	0.798	194
Housing	10201	0.805	133	102	0.844	343
Neighbourhood	10202	0.895	41	102	0.844	343
Travel and transport	10203	0.831	63	102	0.844	343
Environmental exposure	10204	0.738	71	102	0.844	343
Cardiovascular system	10301	0.727	80	103	0.886	1298
Musculoskeletal system	10302	0.697	76	103	0.886	1298
Nervous system	10304	0.771	45	103	0.886	1298
Digestive system	10305	0.836	57	103	0.886	1298
Urogenital system	10306	0.700	51	103	0.886	1298
Endocrine system	10307	0.471	15	103	0.886	1298
Hemic and immune systems	10308	0.222	10	103	0.886	1298
Hearing, vision, speech	10309	0.833	129	103	0.886	1298
Oral/dental health	10310	0.679	51	103	0.886	1298

Congenital malformations	10312	0.625	11	103	0.886	1298
Cancer	10313	0.333	4	103	0.886	1298
Mortality	10314	0.621	17	103	0.886	1298
Women's health	10316	0.854	42	103	0.886	1298
Accidents and injuries	10317	0.894	80	103	0.886	1298
Allergies	10318	0.791	67	103	0.886	1298
Infections	10319	0.746	33	103	0.886	1298
Anthropometry	10320	0.860	98	103	0.886	1298
Physical characteristics	10321	0.764	40	103	0.886	1298
Physical functioning	10322	0.420	49	103	0.886	1298
General health	10323	0.435	132	103	0.886	1298
Mental disorders	10401	0.813	31	104	0.905	906
Personality | Temperament	10402	0.808	226	104	0.905	906
Wellbeing	10403	0.651	49	104	0.905	906
Emotions	10404	0.692	42	104	0.905	906
Cognitive function	10405	0.703	37	104	0.905	906
Health services utilisation	10501	0.508	64	105	0.709	242
Hospital admissions	10502	0.785	69	105	0.709	242
Immunisations	10503	0.556	10	105	0.709	242
Medications	10504	0.750	34	105	0.709	242
Complementary therapies	10505	0.857	12	105	0.709	242
Diet and nutrition	10601	0.898	240	106	0.913	527
Physical activity	10602	0.705	54	106	0.913	527
Alcohol consumption	10605	0.899	80	106	0.913	527
Substance abuse	10606	0.981	52	106	0.913	527
Criminal behaviour	10608	1.000	5	106	0.913	527
Home life	10701	0.740	77	107	0.875	786
Household composition	10702	0.764	80	107	0.875	786
Marital status	10703	0.750	65	107	0.875	786
Family members and relations	10704	0.728	141	107	0.875	786
Friends	10705	0.560	28	107	0.875	786
Childcare	10706	0.679	27	107	0.875	786
Child welfare	10707	0.667	9	107	0.875	786
Social support	10708	0.765	99	107	0.875	786
Leisure activities	10709	0.749	90	107	0.875	786
Technology	10711	0.706	17	107	0.875	786
Qualifications	10801	0.937	88	108	0.855	588
Further education | Higher education	10803	0.676	38	108	0.855	588
Training	10804	0.621	25	108	0.855	588
Basic skills	10805	0.722	56	108	0.855	588

Adult education	10806	0.667	7	108	0.855	588
Learning difficulties	10807	0.717	28	108	0.855	588
Pre-school	10808	0.400	6	108	0.855	588
Cognitive skills	10810	0.759	16	108	0.855	588
Non cognitive skills	10811	0.200	5	108	0.855	588
Education aspirations	10813	0.500	15	108	0.855	588
Primary schooling	10815	0.756	20	108	0.855	588
Occupation | Employment	10901	0.857	373	109	0.887	675
Social classification	10902	0.846	15	109	0.887	675
Income	10903	0.750	61	109	0.887	675
Finances	10904	0.766	56	109	0.887	675
Assets	10905	0.500	5	109	0.887	675
Consumption | Expenditure	10906	0.714	24	109	0.887	675
Pensions	10907	0.500	3	109	0.887	675
Benefits | Welfare	10908	0.833	13	109	0.887	675
Social attitudes	11001	0.600	5	110	0.670	121
Politics	11002	0.800	19	110	0.670	121
Infant feeding	11101	0.703	33	111	0.872	562
Language and vocabulary	11102	0.907	80	111	0.872	562
Parenting	11103	0.816	177	111	0.872	562
Developmental milestones	11104	0.889	60	111	0.872	562
Childbirth	11401	0.763	33	114	0.723	116
Macro average		**0.724**	6736		**0.830**	7034
Weighted average		**0.787**	6736		**0.866**	7034

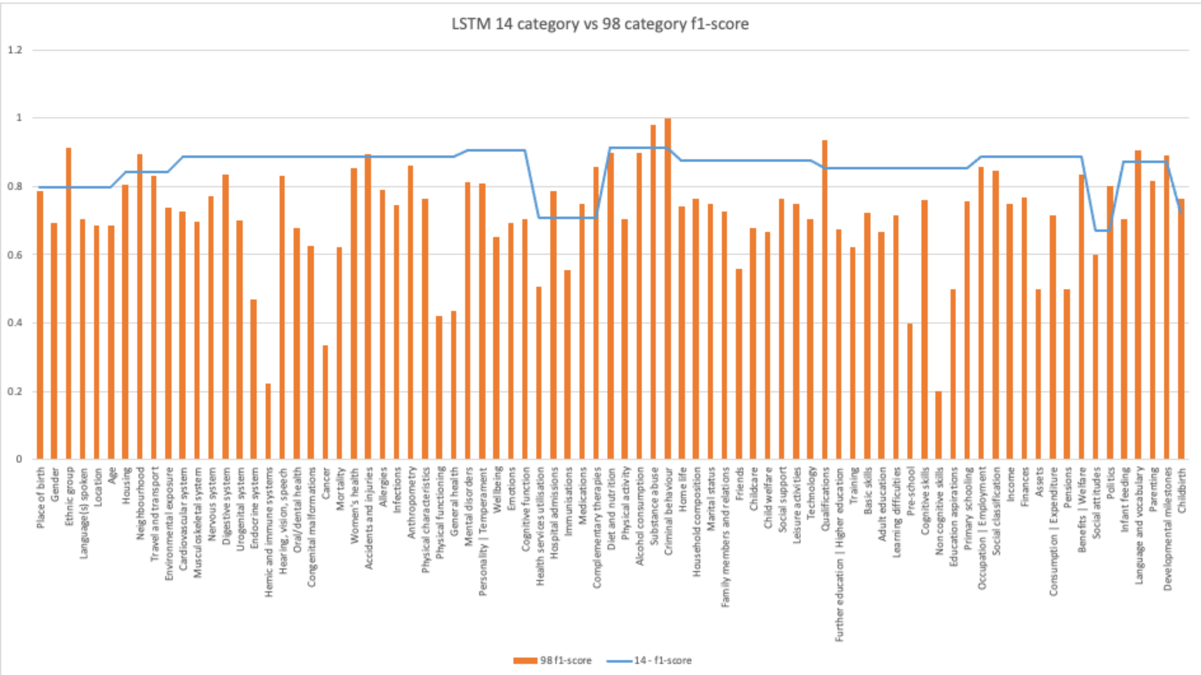


Fig.4. Per-class F1-scores for the Simple LSTM model in both 14-class and 98-class classification.

Table 4: Second level per-class and aggregate (in bold) f1-scores for the Multinomial Naive Bayes model trained with question-response concatenation and the addition of section heading metadata.

Category	98 CODE	98 f1-score	support	14 CODE	14 - f1-score	dataset size
Place of birth	10101	0.000	15	101	0.663	203
Gender	10102	0.500	18	101	0.663	203
Ethnic group	10103	0.000	21	101	0.663	203
Language(s) spoken	10104	0.000	10	101	0.663	203
Location	10106	0.000	21	101	0.663	203
Age	10107	0.000	20	101	0.663	203
Housing	10201	0.718	137	102	0.776	355
Neighbourhood	10202	0.542	43	102	0.776	355
Travel and transport	10203	0.611	64	102	0.776	355
Environmental exposure	10204	0.632	77	102	0.776	355
Cardiovascular system	10301	0.496	83	103	0.827	1372
Musculoskeletal system	10302	0.400	81	103	0.827	1372
Nervous system	10304	0.296	46	103	0.827	1372
Digestive system	10305	0.583	60	103	0.827	1372
Urogenital system	10306	0.632	52	103	0.827	1372
Endocrine system	10307	0.000	16	103	0.827	1372
Hemic and immune systems	10308	0.000	10	103	0.827	1372
Hearing, vision, speech	10309	0.773	137	103	0.827	1372
Oral/dental health	10310	0.290	52	103	0.827	1372
Congenital malformations	10312	0.000	10	103	0.827	1372
Cancer	10313	0.000	4	103	0.827	1372
Mortality	10314	0.000	17	103	0.827	1372
Women's health	10316	0.275	44	103	0.827	1372
Accidents and injuries	10317	0.781	82	103	0.827	1372
Allergies	10318	0.765	70	103	0.827	1372
Infections	10319	0.553	34	103	0.827	1372
Anthropometry	10320	0.820	100	103	0.827	1372
Physical characteristics	10321	0.586	41	103	0.827	1372
Physical functioning	10322	0.382	54	103	0.827	1372
General health	10323	0.316	138	103	0.827	1372
Mental disorders	10401	0.000	32	104	0.839	942
Personality | Temperament	10402	0.705	238	104	0.839	942
Wellbeing	10403	0.207	51	104	0.839	942
Emotions	10404	0.000	47	104	0.839	942
Cognitive function	10405	0.093	41	104	0.839	942
Health services utilisation	10501	0.192	64	105	0.459	251
Hospital admissions	10502	0.491	72	105	0.459	251

Immunisations	10503	0.000	10	105	0.459	251
Medications	10504	0.449	38	105	0.459	251
Complementary therapies	10505	0.533	11	105	0.459	251
Diet and nutrition	10601	0.865	255	106	0.886	542
Physical activity	10602	0.459	56	106	0.886	542
Alcohol consumption	10605	0.786	84	106	0.886	542
Substance abuse	10606	0.835	52	106	0.886	542
Criminal behaviour	10608	0.000	5	106	0.886	542
Home life	10701	0.542	78	107	0.803	823
Household composition	10702	0.569	83	107	0.803	823
Marital status	10703	0.654	67	107	0.803	823
Family members and relations	10704	0.653	152	107	0.803	823
Friends	10705	0.000	29	107	0.803	823
Childcare	10706	0.000	27	107	0.803	823
Child welfare	10707	0.000	9	107	0.803	823
Social support	10708	0.654	104	107	0.803	823
Leisure activities	10709	0.544	93	107	0.803	823
Technology	10711	0.000	17	107	0.803	823
Qualifications	10801	0.837	95	108	0.828	617
Further education | Higher education	10803	0.273	38	108	0.828	617
Training	10804	0.000	25	108	0.828	617
Basic skills	10805	0.351	57	108	0.828	617
Adult education	10806	0.000	7	108	0.828	617
Learning difficulties	10807	0.000	28	108	0.828	617
Pre-school	10808	0.000	6	108	0.828	617
Cognitive skills	10810	0.000	17	108	0.828	617
Non cognitive skills	10811	0.000	5	108	0.828	617
Education aspirations	10813	0.000	15	108	0.828	617
Primary schooling	10815	0.483	22	108	0.828	617
Occupation | Employment	10901	0.514	385	109	0.864	711
Social classification	10902	0.000	15	109	0.864	711
Income	10903	0.638	62	109	0.864	711
Finances	10904	0.417	57	109	0.864	711
Assets	10905	0.000	5	109	0.864	711
Consumption | Expenditure	10906	0.000	23	109	0.864	711
Pensions	10907	0.000	2	109	0.864	711
Benefits | Welfare	10908	0.556	13	109	0.864	711
Social attitudes	11001	0.000	5	110	0.353	126
Politics	11002	0.273	19	110	0.353	126
Infant feeding	11101	0.211	34	111	0.796	585
Language and vocabulary	11102	0.588	83	111	0.796	585
Parenting	11103	0.734	186	111	0.796	585

Developmental milestones	11104	0.692	64	111	0.796	585
Childbirth	11401	0.522	34	114	0.261	119
Macro average		**0.370**	7032		**0.702**	7034
Weighted average		**0.556**	7032		**0.789**	7034

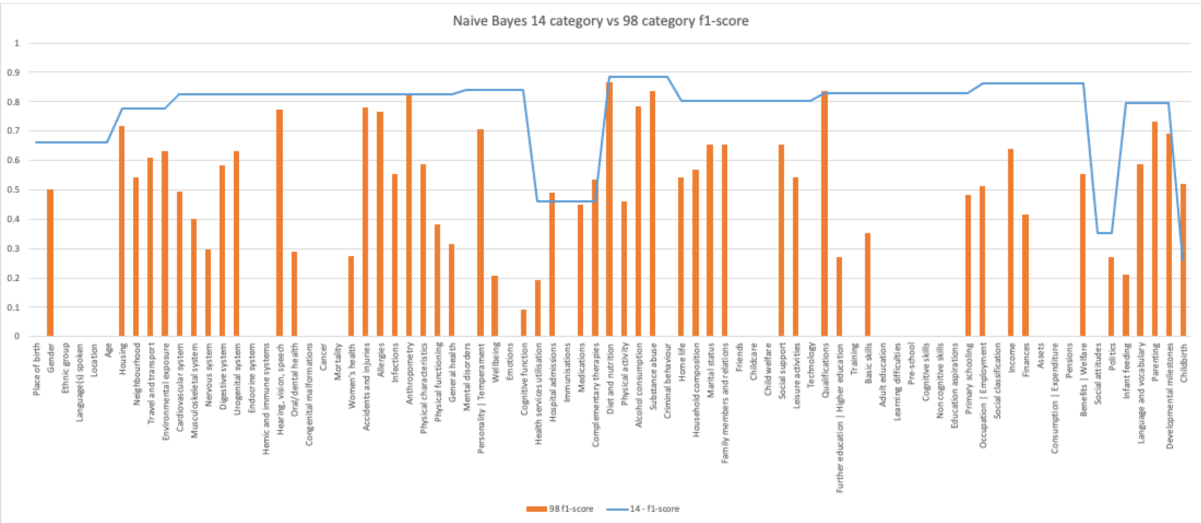


Fig.5. Per-class F1-scores for the Multinomial Naive Bayes model in both 14-class and 98-class classification.

The maximum weighted average F1 score in the 14-class problem is achieved by the ULMFit model, closely followed by the BERT_base_uncased model. The Multinomial Naive Bayes model is used as a baseline to rule out any neural network-based models that fall under the performance of this less computationally demanding model. BERT_base_uncased achieves the maximum weighted F1 score in the 98-class classification task and appears to produce more stable performance across the two tasks than the ULMFit model.

2. Concept prediction - evaluation against a range of different types of unseen data

Previously unexplored dimensions of the training dataset are its representativeness for new data (in this case unseen studies) and whether changes in the way similar questions (by vocabulary category) are asked in different domains (social science vs biomedical) ask the same questions.

We take questions annotated to the CLOSER vocabulary, from new studies from a social science and biomedical domain, remove the annotation and examine the F1 score (prediction) evaluation metric results with that manually tagged.

In order to see how the trained models work on a new dataset, we used “Health and Employment After Fifty” (HEAF) (study website: https://www.mrc.soton.ac.uk/heaf/) questionnaires as the unseen data. The data was obtained the same way as the training data, i.e. using the API from Closer Discovery (described in section 3). The models were then run in inference mode on this new data.

Table 5: Aggregate and per-category f1-scores for all considered models derived from inference using the HEAF dataset. All models were trained with features using question-response concatenation and section-heading metadata.

Category	BERT base uncased	Multinomial Naive Bayes	ULMFit	Simple LSTM	Number of items
Administration	0.000	0.000	0.000	0.000	0
Child development	0.000	0.000	0.000	0.000	0
Demographics	0.500	1.000	1.000	0.727	6
Education	0.000	0.000	0.000	0.000	0
Employment and income	0.889	0.765	0.722	0.837	191
Expectations, attitudes and beliefs	0.000	0.000	0.000	0.000	0
Family and social networks	0.533	0.467	0.316	0.737	7
Health behaviour	0.837	0.833	0.810	0.718	21
Health care	0.000	0.000	0.000	0.000	0
Housing and local environment	0.000	0.000	0.000	0.000	0
Life events	0.000	0.000	0.000	0.000	14
Mental health and mental processes	0.000	0.000	0.000	0.000	6
Physical health	0.862	0.708	0.682	0.721	53
Pregnancy	0.000	0.000	0.000	0.000	0
Micro average	0.805	0.691	0.628	0.715	298
Macro average	0.259	0.270	0.252	0.267	298
Weighted average	**0.805**	**0.706**	**0.668**	**0.747**	298

Table 5 shows the per-class and aggregate F1 scores for each considered model with the HEAF dataset. The *micro average* F1 score is calculated using a precision and recall taken from the total precision and total recall summed over all samples, and does not consider category size. Again, all models shown represent the variant trained on data question-response concatenation and the addition of section heading metadata. It is clear from the weighted averages in table 5 that the BERT-type model represents by some distance the greatest performance over this unseen dataset, and may be the preferred model for further work with unseen datasets. The *generalisability* of a trained neural network model, the ability to retain model performance seen in test data taken from the training data corpus and novel unseen data, is one of the clearest indicators of ultimate model utility. The high F1 score seen in the BERT_base_uncased model on this unseen dataset suggests a low-level of *overtraining* on the initial training dataset, although assessment of performance on additional unseen datasets will be required to determine this conclusively.

3. Concept prediction - understanding the relationship between training dataset size and prediction, i.e. the minimum training dataset set size and compositions

Previous work (Fig. 1) has established that F1 score by model varies across different vocabulary terms. Understanding the inflexion point where the F1 score drops will provide a deeper understanding of this relationship, providing guidance for the development of further training datasets for concept prediction in other languages and vocabularies. The full CLOSER vocabulary contains > 120 categories, so this will assist in identifying areas where the composition of the training datasets could be improved to get equitable prediction across all potential vocabulary categories.

The training dataset was extracted using Python 3 code (Li, J., 2021) from CLOSER Discovery utilising the Colectica Repository REST API (Colectica, 2021). The training dataset generation is described in (De et al., 2022).

The training dataset contained approx. 36000 rows, composed of question text, response domain and annotated with the 16 item CLOSER vocabulary (https://wiki.ucl.ac.uk/display/CLOS/Topics).

The dataset was randomly segmented into deciles of decreasing size. Multinomial Naive Bayes was taken to be the baseline model, against which all other considered models were compared. Aggregate F1-scores for each model at each dataset size are provided in table 6.

Table 6. F1-score by dataset size and vocabulary category. F1-scores are evaluated against the Multinomial Naive Bayes score at the full dataset size of 0.702 (macro average) and 0.789 (weighted average). Entries are coloured red if they are below this baseline and green if they are equal to or above. For each dataset size, the largest f1-score is shown in bold.

Measure	Dataset size	BERT F1	ULMFit F1	Simple LSTM F1	Naive Bayes F1
macro avg	10	0.249	**0.523**	0.291	0.440
macro avg	20	0.499	**0.659**	0.517	0.515
macro avg	30	0.631	**0.748**	0.635	0.564
macro avg	40	0.721	**0.794**	0.674	0.586
macro avg	50	0.778	**0.803**	0.705	0.649
macro avg	60	0.781	**0.833**	0.737	0.655

macro avg	70	0.805	**0.859**	0.780	0.678
macro avg	80	0.813	**0.853**	0.798	0.669
macro avg	90	0.808	**0.863**	0.817	0.685
weighted avg	10	0.421	**0.592**	0.389	0.551
weighted avg	20	0.656	**0.718**	0.616	0.653
weighted avg	30	0.748	**0.787**	0.709	0.693
weighted avg	40	0.799	**0.832**	0.741	0.705
weighted avg	50	0.828	**0.837**	0.775	0.752
weighted avg	60	0.837	**0.860**	0.794	0.758
weighted avg	70	0.848	**0.880**	0.828	0.771
weighted avg	80	0.860	**0.878**	0.837	0.771
weighted avg	90	0.859	**0.888**	0.859	0.780

From these results it is clear that ULMFit is the most robust neural-network based model across different dataset sizes, followed closely by BERT_base_uncased. The low macro averaged f1-scores at low dataset size show that BERT may be more substantially affected by under-represented classes when compared with ULMFit, although BERT suffers less with respect to ULMFit when considering the weighted average f1-score.

4. Concept prediction - investigating hierarchical and multi-label approaches for second-level topic classification

Hierarchical classification approaches for the second-level topic classification task, for classifying a question text into its relevant top-level and second-level hierarchy, have been investigated through deep neural networks. Hierarchical document classifiers have been designed using a Recurrent Neural Networks (RNN) to implement the layered structure of nonlinear processing components. The developed approach considers the entire training dataset in the first step of the top-level topic classification. The second level of prediction is done by lowering and narrowing the next set of inputs as the child nodes from the output of the top-level prediction. These are then extended to incorporate an attention layer to emphasise distinct areas of the text's semantic representation.

An attention function is generally used to describe the mapping of a query and a collection of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is a weighted sum of the values, with each value assigned a weight defined by the query's compatibility function with the relevant key. Here the attention mechanism focuses on phrase segments, with the significance of the segment defined by its contribution to the job.

In the attention mechanism, by combining the encoder output and decoder output at timestamp t, a context vector is created. The encoder's most relevant information is included into the context vector. Following data pre-processing (punctuation removal, lowercasing and conversion of 5-digit encoded top and second-level category into separated top-level and second-level columns in the dataset), the top-level and second-levels (level 1 and level 2 for hierarchical classification) are converted into a dictionary with the appropriate key and value pair. The model architecture consists of a Gated

Recurrent unit (GRU) with 100 cells and a dropout percentage of 0.2%. The GRU sequential model is supplemented by a GloVe embedding layer that uses the ‘n’ unique words in our dataset, which totals 9109 tokens.

For each input sentence, a sequence of annotations are generated by the Bidirectional GRU. These vectors are obtained by concatenation of forward and backward hidden states in the encoder, with the context vector constructed by concentrating on the word embeddings in the input that are represented by hidden states, and this is accomplished by simply adding the weighted sums of the hidden states together. The loss function used is ‘sparse_categorical_crossentropy’ with a Softmax activation function and RMSProp as the optimiser over 10 epochs. Figure 6 shows the per-class f1-score for the second-level 100-class problem for a sample of representative classes (each second-level topic is specified with the fully qualified top level-second level naming convention).

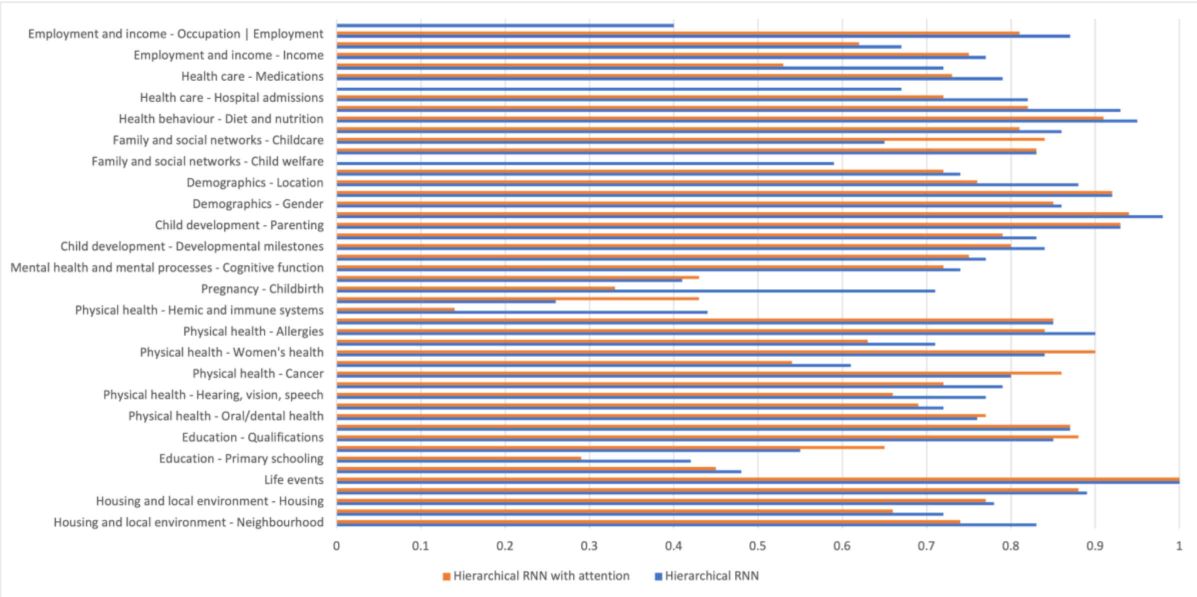


Fig.6. Per-class F1-scores for RNN and RNN-with-attention hierarchical models for second-level topic classification.

From the figure, hiRNN gives a slightly better performance than with the attention mechanism, for some of the second-level topics, though the difference is not appreciable in many cases. This is more prominent in cases where the question text is very short. The next planned method is to investigate ‘teacher forcing’ to train an RNN model such as LSTM. Instead of using the last generated word as the next input to the decoder, teacher forcing uses the target word and the loss is recorded. This avoids very poor results during the early stages of training as the decoder is being corrected, enabling the training to converge much quicker.

5. Conclusions and Future Work

In this work-package, we have investigated the impact on question text topic/category classification from various aspects, namely,

- type of ML model architecture,
- size of the training dataset
- level of heterogeneity in the composition of the dataset (14-class versus 98-class)

It is evident that the difference in prediction performance in the top-level (14-class) versus the second-level of topics (98-class) can be attributed to both the size and composition of the supporting training samples. While neural network-based ML models deliver improved performance with bigger training sets, the CLOSER dataset also has the additional challenge of label bias in the annotation of the top-level and second-level topics (where the annotation of question texts to specific topics from the CLOSER ontology is performed by the corresponding experts who performed the study) as well as the influence of semantic divergence within each top-level and second-level topics. Intuitively, the level of semantic divergence is higher at the 14-class level. This challenge is not particular to the particular CLOSER ontology, but points to the need for measures to be put into place for a training dataset to achieve a given level of prediction. A direction of investigation in this regard is the semantic heterogeneity within different topic levels, where topic modelling followed by dimensionality reduction (to cluster question texts with similar semantics) is a promising approach.

An important finding from the work to date is the apparently significant, though clearly noticeable difference, contextual metadata makes to the level of prediction. Further evaluation and quantification of this will be an important outcome for future work, as it could have a large impact on both the size and complexity of and the time, effort and cost implied if trying to construct training datasets to support other ontologies.

Another subsequent planned step (beyond the objectives of this work-package) is to investigate the questionnaire structure and its influence on classification performance. This is elaborated in the following sub-section.

Dependency modelling and Sequence Labelling

Given the nature of the task in hand, it is safe to assume that the questions within the same questionnaire follow a similar theme. Here, we are investigating whether there is an interdependence between the chain of questions within the same questionnaire. Specifically, to simplify the problem we first reduce this investigation to a sequence of two questions, i.e. each question and its followup.

To conduct this analysis, at both the super and sub level categories, we will simply look at the statistics that highlight this dependency.

What each of these charts show, is the frequency of each category following another category. What the charts show is that (varingly) for each category there is a dominant pre-category. This can serve as the basis of the depency that we would like to encode and add to our existing deep learning based model.

In the pie charts below, we see the distribution of the previous categories given each category.

As it can be seen apart from a few exceptions, in almost all categories, the preceding subcategory is identical. The same is true for lower level 96 categories. Here are the charts for the first 15.

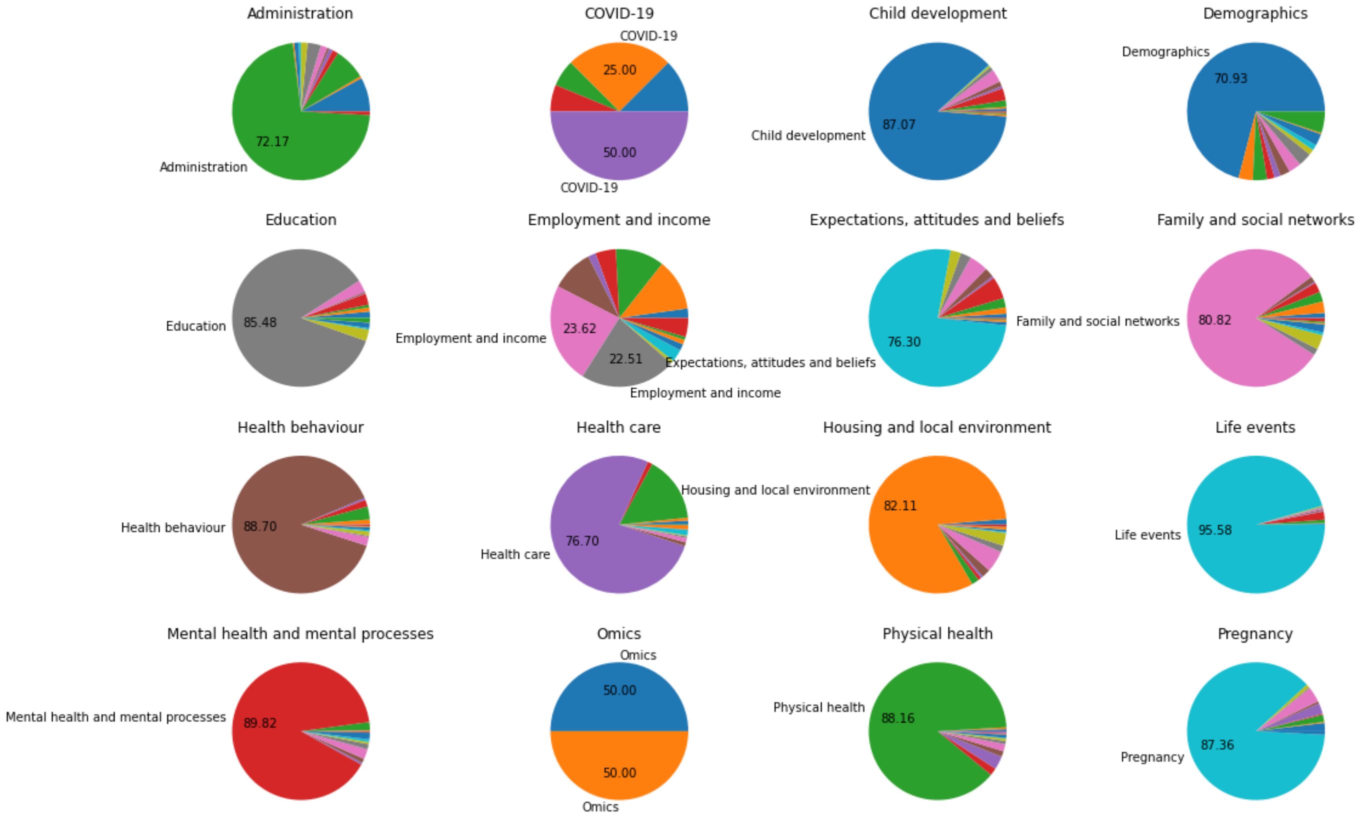


Fig.7. Previous category visualisation for the 16 top-level categories.

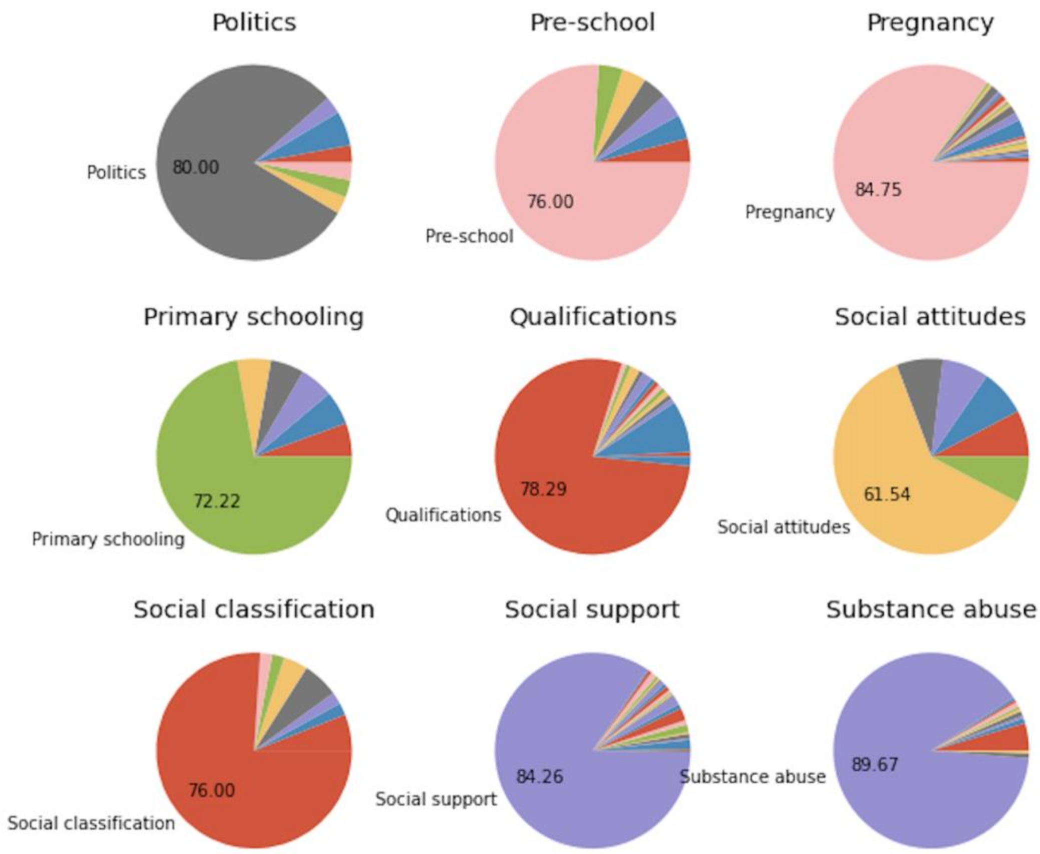


Fig.8. Previous category visualisation for the second-level categories; showing a dominant previous category

In almost all subcategories, we had strong dependency between two consequent tags, where both were identical between 96 sub-categories. The charts above are a small selection of these charts, but the trend holds for other categories. There are a few exceptions in subcategories such as:

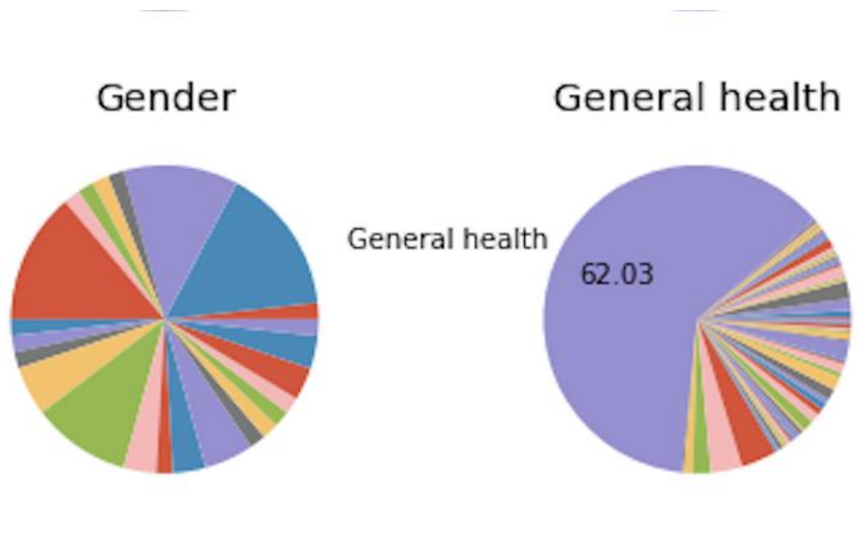
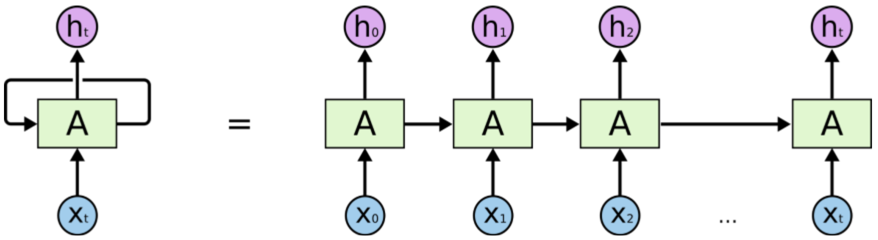


Fig.9. Previous category visualisation for the second-level categories; with non-dominant previous categories.

Given the strong dependency that is evident here between two consequent labels, for future work we would like to design models where their architecture allows us to apply conditional training to learn **dependencies** among consequent labels.
We will investigate one of the following approaches that are common in sequence labelling problems in NLP:

1) Classical Approaches: mostly rule-based. where we manually devise heuristics and code them.
2) Classical Machine Learning Approaches: Models such as Conditional Random Field (CRF). It is a probabilistic graphical model that can be used to model sequential data such as labels of words in a sentence. The CRF model is able to capture the features of the current and previous labels in a sequence but it cannot understand the context of the forward labels.
3) Deep Learning Approaches:
    a) Recurrent neural networks (RNN) are a class of deep neural networks that are powerful for modelling sequence data such as time series, or natural languageAs described in Keras/tensorflow guide for RNNS, “Schematically, a RNN layer uses a for loop to iterate over the timesteps of a sequence, while maintaining an internal state that encodes information about the timesteps it has seen so far.”



    This chain-like nature reveals that recurrent neural networks are intimately related to sequences. They are the natural architecture of neural networks to use for such data.

    b) Long short Term Memory (LSTM). We plan to investigate the use of bi-directional LSTMs because using a standard LSTM to make predictions will only take the “past” information in a sequence into account. Two different state-of-the-art LSTM architectures that can be applicable to our problem are:

        i) Bidirectional LSTM-CRF:
        For sequence tagging with Bidirectional LSTM-CRT, please refer to this implementation in keras.

        ii) Bidirectional LSTM-CNNs:
        More details and implementation in keras.

References

Colectica (2021). Colectica. Available at: https://www.colectica.com. Accessed 4 Oct. 2021.

CLOSER Discovery (2021). Available at:https://discovery.closer.ac.uk/. Accessed 30 March 2022.

De, S., Moss, H., Johnson, J., Li, J., Pereira, H., & Jabbari, S. (2022). Engineering a machine learning pipeline for automating metadata extraction from longitudinal survey questionnaires. IASSIST Quarterly, 46(1).

Devlin, J., et al. BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

Goyal, A., Lamb, A., Zhang, Y., Zhang, S., Courville, A. & Bengio, Y. (2016), 'Professor forcing: A new algorithm for training recurrent networks', Advances in Neural Information Processing Systems pp. 4608–4616. URL: https://arxiv.org/abs/1610.09038v1

Howard, J, and Ruder, S., Universal language model fine-tuning for text classification (2018), arXiv preprint arXiv:1801.06146.

Kuprieiev, R. et al. (2021). *DVC: Data Version Control - Git for Data & Models (2.3.0)* DOI:10.5281/zenodo.4892897

Li, J. (2021). Python interface to the Colectica API (Version 1.0). Computer software. https://github.com/CLOSER-Cohorts/colectica_api. Accessed 4 October 2021.

Merity, S., Keskar, N. S., and Socher, S. Regularizing and optimizing LSTM language models (2017), arXiv preprint arXiv:1708.02182.

UCL (2021). Research Data Storage Service (Research software repository). Available at https://www.ucl.ac.uk/isd/services/research-it/research-data-storage-service. Accessed 14 Oct. 2021.