Social Computing Group,
Department of Informatics
Technical University of Munich

Explaining Neural NLP Models for the Joint Analysis of Open- and Closed-Ended Survey Answers

Edoardo Mosca, Katharina Hermann, Tobias Eder, Georg Groh

Motivation and Objectives

- Surveys are **a popular tool to collect data** (scientific studies, census questionnaires, customer feedback)
- **Open-** and **closed-ended** answers provide the most insights when **combined**.
- Previous works employ human labour or shallow machine learning models. We investigate the usage of **NLP transformers + XAI techniques**.

Engineering Major Survey (EMS)

From **2015** to **2019**

7197 surveyed students from **27** US universities

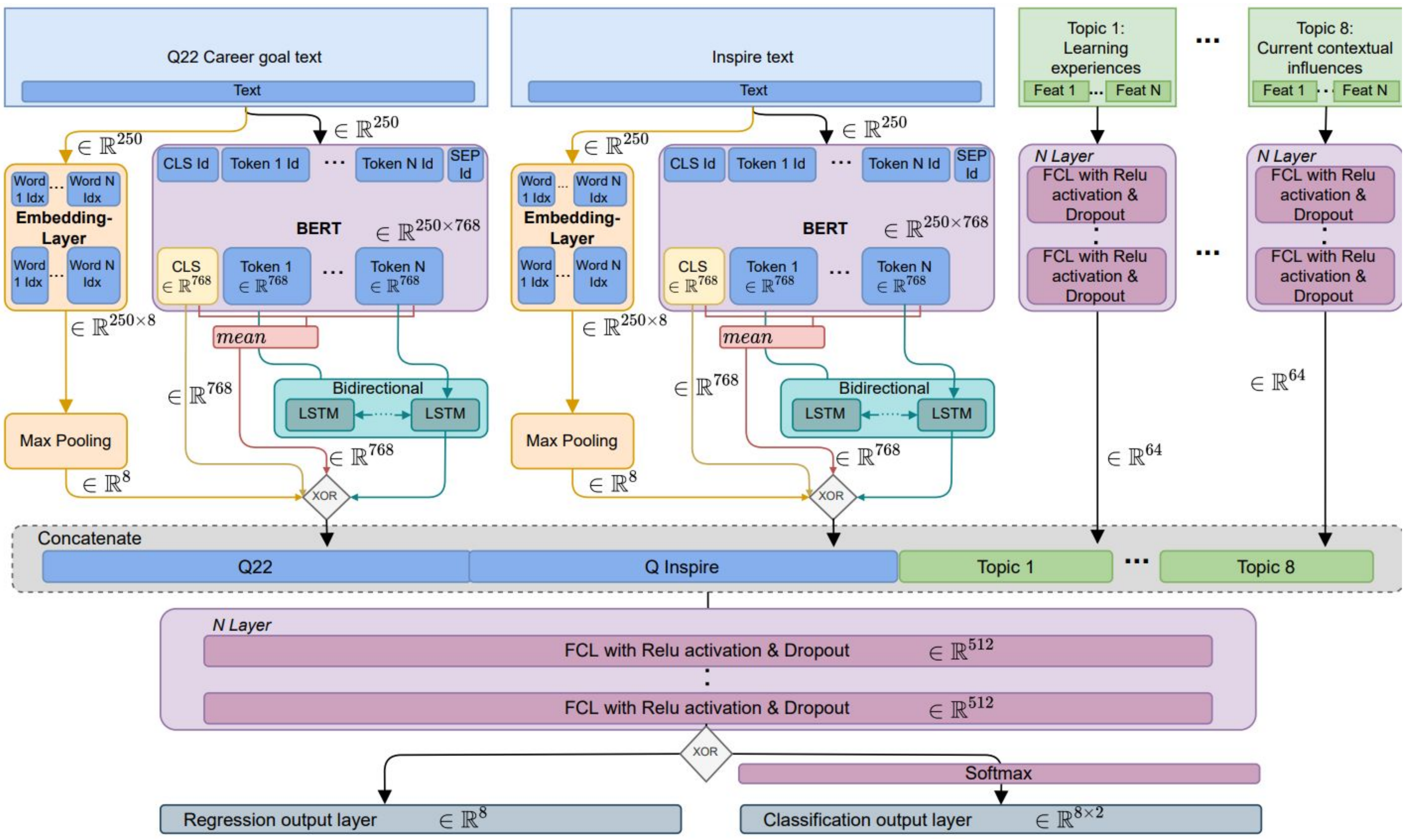
Longitudinal study of college students. Studies how factors from specific topics + open text variables influence their desired career path.

Topic 1: Learning experiences.. **topic 5**: Background.. **topic 8**: Current contextual influences.

+

Q22: *"[…] If you would like to elaborate on what you are planning to do, in the next five years or beyond, please do so here."*
Inspire: *"To what extent did this survey inspire you to think about your education in new or different ways? Please describe."*

T1: Work for a small business / start-up
T2: Work for a medium/large company
T3: Work for a non-profit organization
T4: Work for the government, military, or public agency.
T5: Work as a teacher in a K-12 school
T6: Work as a faculty member in a college/university
T7: Found your own for-profit organization
T8: Found your own non-profit organization



Multi-modal model:

- BERT for open-ended answers
- FCLs for closed-ended questions.

Ablation study:

XORs indicate different architecture choices.

Explaining the model:

We apply SHAP and ConceptSHAP at several model levels to get a holistic understanding.

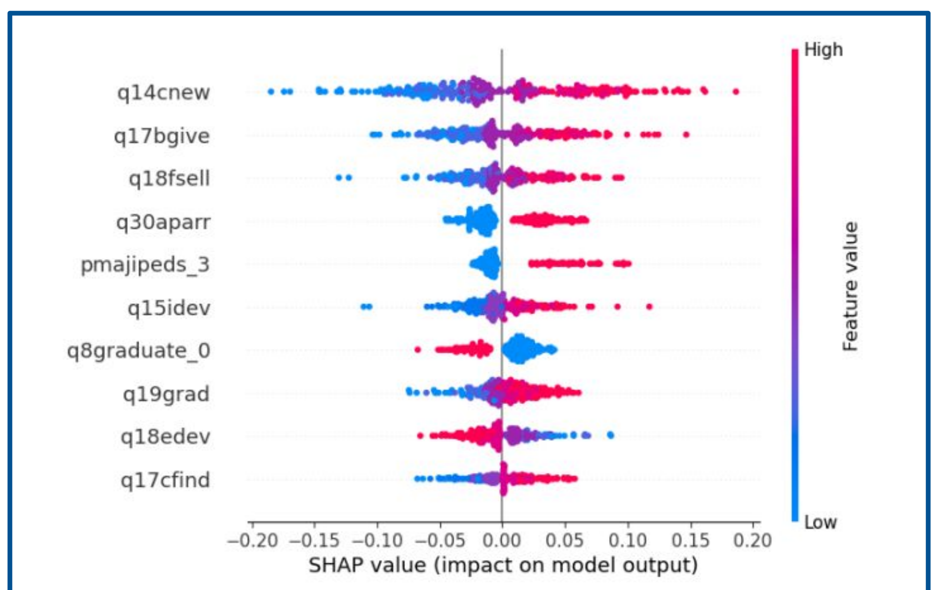
Task Results

Architecture			T1	T2	T3	T4	T5	T6	T7	T8
Q22	no T	C	51.66	60.10	56.89	44.61	48.40	51.85	52.50	63.70
		R	53.82	51.36	50.82	58.75	43.63	42.24	46.71	62.40
Ins.	no T	C	46.66	38.20	40.68	42.20	50.21	43.48	46.08	42.69
		R	42.26	39.79	36.07	37.77	37.10	41.79	41.88	35.48
Q22+Ins.	no T	C	45.69	59.87	52.31	53.11	47.92	59.71	50.91	51.12
		R	**63.48**	47.46	50.59	45.20	41.06	41.29	39.86	58.73
No text	all T	C	50.85	53.34	61.03	52.40	57.03	**67.88**	61.02	72.65
		R	50.79	54.17	61.58	57.33	**58.94**	56.91	59.08	74.65
Q22	all T	C	63.01	60.74	**63.53**	**60.87**	50.77	57.76	54.90	73.64
		R	59.69	**63.64**	59.59	55.84	56.62	56.03	**62.66**	**76.23**
Ins.	all T	C	57.23	59.08	57.63	54.22	54.68	57.48	65.30	69.24
		R	48.33	47.00	51.49	50.45	48.92	46.12	58.49	72.47
Q22+Ins.	all T	C	58.71	57.52	59.86	55.51	55.16	58.56	62.40	71.55
		R	59.49	54.62	63.27	55.50	56.83	49.58	56.60	73.61

Simple aggregation of BERT embeddings works best.

	CLS	Mean	BiLSTM	Embedding
C	60.66	**63.70**	37.88	49.66
R	53.96	62.40	58.18	50.27

Model Explanations



(a) Global expl., embedding and numerical feature inputs

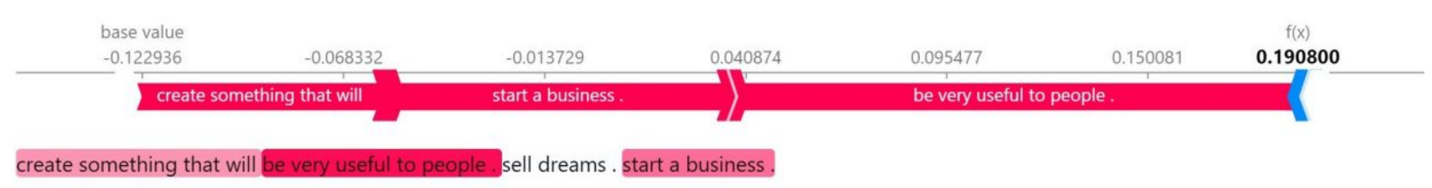
Lack of specific direction / broad plans

Clear path, self-centeredness , determination

Concepts Learned

Plan type, desired career / work place

Time planning, distance from goal



(a) Local explanation: text relevance w.r.t. specific neuron

(b) Local explanation: text relevance w.r.t. model output

Takeaways

- Multi-modal models can **leverage transformers** to **analyze jointly** open- and closed-ended answers.
- More **scalable** and **accurate** than previous solutions. Qualitative results are in-line with what extracted manually by human analysts.
- Combining **feature-attribution** and **concept explanations** provides us with a **holistic understanding** of what the model has learnt.

Check out our code !