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

Edoardo Mosca, Katharina Hermann, Tobias Eder and **Georg Groh**
TU Munich, Department of Informatics, Germany
{edoardo.mosca, katharina.hermann, tobi.eder}@tum.de
grohg@in.tum.de

Abstract

Large-scale surveys are a widely used instrument to collect data from a target audience. Beyond the single individual, an appropriate analysis of the answers can reveal trends and patterns and thus generate new insights and knowledge for researchers. Current analysis practices employ shallow machine learning methods or rely on (biased) human judgment. This work investigates the usage of state-of-the-art NLP models such as BERT to automatically extract information from both open- and closed-ended questions. We also leverage explainability methods at different levels of granularity to further derive knowledge from the analysis model. Experiments on EMS—a survey-based study researching influencing factors affecting a student's career goals—show that the proposed approach can identify such factors both at the input- and higher concept-level.

1 Introduction

Surveys and questionnaires are prevalent tools to inquire about an audience and collect ideas, opinions, and thoughts. Common examples are requesting user feedback concerning a specific product or service, regular reports for scientific studies that involve human subjects, and census questionnaires directed to a certain demographic population.

Carrying out an appropriate and thorough analysis of the collected answers is of major relevance for researchers both in the industry and academia. However, the generated data are often a combination of open-ended and closed-ended questions. While the former gathers a participant's thoughts in text form, the latter consists in selecting one (or more) of the options specified by the survey designer. Utilizing both types remains a popular choice as closed-ended questions are very suitable to derive statistical conclusions but may lack details which are in turn provided by open-ended answers.

Currently, the two dominant analysis practices comprise traditional closed-vocabulary and open-vocabulary methods (Eichstaedt et al., 2021). Whereas the former introduces human biases and is resource-intensive, the latter overcomes these challenges with the help of *Natural Language Processing* (NLP) techniques. Nonetheless, both approaches fail to consider contextual information and do not leverage currently available NLP architectures to deal with more complex patterns.

In this work, we bridge the gap in research and investigate the usage of deep-learning-based methods from NLP and explainability techniques to extract knowledge and interpret correlations from surveys presenting both structured and unstructured components. Our contribution can be summarized as follows:

(1) We apply a popular transformer architecture (DistilBERT) (Sanh et al., 2019) to open-ended questions. This enables our approach to extract contextual correlations from the text with high precision compared to traditional methods.

(2) Due to the model's black-box characteristics, we utilize post-hoc explainability methods to interpret the extracted correlations. Specifically, we utilize several variants of *SHapley Additive exPlanations* (SHAP) (Lundberg and Lee, 2017) to analyze both instance-level feature importance as well as high-level concepts learned by the model (Yeh et al., 2020). These methods are applied to several components to generate a holistic understanding of the model used for the analysis.

(3) Our approach delivers promising results on the EMS 1.0 dataset - studying influencing factors in students' career goals (Gilmartin et al., 2017). First, it identifies the most relevant factors from closed-ended responses with high precision. Second, it also automatically reveals influencing factors from the open-ended text answers.

Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 49 - 63

2 Related Work

2.1 The EMS Study and Entrepreneurial Behavior Predictors

In this paper, we work with the *Engineering Major Survey* (EMS) longitudinal study of students' career goals by Gilmartin et al. (2017). Analysis of the contents of this study was previously conducted mainly by the social sciences with a focus on qualitative approaches to extract the most influential variables on career goals (Grau et al., 2016; Levine et al., 2017). Quantitative correlation between variables was previously explored by Atwood et al. (2020) relating *Social Cognitive Career Theory* (SCCT) (Lent et al., 1994) to different predefined topics for the purpose of survey design, such as students demographics, first-generation status, and family background. Schar et al. (2017) meanwhile focused on the variables *Engineering Task Self-Efficacy* and *Innovation Self-Efficacy* through explainable regression models.

2.2 Analysis of Open-ended Survey Question in the Social Sciences

In the social sciences, textual analysis has a long history of utilizing manual analysis methods such as *Grounded Theory Method* (GMT) Bryant and Charmaz (2007). However recently, automated text analysis has been used for both open- and closed-vocabulary methods.

Closed-vocabulary methods: Analysis is done by working with a hand-crafted closed-vocabulary such as LIWC (Pennebaker et al., 2001) and calculating the relative frequencies of dictionaries with respect to the text (Eichstaedt et al., 2021).

Open-vocabulary methods: Following the GMT method, these approaches aim to discover topics from data, rather than from a predefined word list (Roberts et al., 2014). For instance, Guetterman et al. (2018) uses NLP techniques such as topic modeling and clustering for textual analysis of survey questions. These approaches were mostly utilizing well-known bag-of-words methods such as *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) and *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990). Further work included clustering semantic distances in adjectives for situation-taxonomies (Parrigon et al., 2017).

2.3 Post-Hoc Explainability

Methods from *eXplainable Artificial Intelligence* (XAI) (Arrieta et al., 2020; Mosca et al., 2021) have recently gained popularity as deep architectures—such as transformers—behave like black-boxes (Brown et al., 2020; Devlin et al., 2019). In particular, post-hoc explainability techniques are able to explain the *why* behind a certain prediction even if the model is not inherently interpretable.

The literature has classified existing interpretability approaches in structured taxonomies depending on their core characteristics (Madsen et al., 2021; Doshi-Velez and Kim, 2017). We identify the following two broad categories as the most relevant for our research objectives and methodology.

Feature attribution methods: They assign each input feature with a relevance score describing its importance for the model prediction. Approaches such as SAGE (Covert et al., 2020) and GAM (Ibrahim et al., 2019) produce global explanations, i.e. at the dataset level. Others, instead, focus on generating insights at the instance-level, i.e. about a specific model prediction. Prominent local methods are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017).

Concept-based methods: Concept-oriented techniques aim at extracting human-interpretable concepts, consisting of sets of (text) features from several input samples sharing similar activation patterns within the model. Prominent approaches are TCAV (Kim et al., 2018), ACE (Ghorbani et al., 2019), and ConceptSHAP (Yeh et al., 2020). The latter is unsupervised—i.e. it does not require a predefined list of concepts to test for—and thus particularly relevant for our methodology.

Please note that these explainability techniques can be applied to the whole model—i.e. from input to output—or sub-components of it, such as (groups of) layers and neurons (Sajjad et al., 2021).

3 Methodology

3.1 EMS Data

We use the EMS 1.0 data as our data source and prediction target. The EMS study 1.0 from 2015 consists of data from 7,197 students enrolled across 27 universities in the United States. The study poses a mix of closed and free-text questions across 8 different topics, ranging from background characteristics to self-efficacy and career goals. More de-

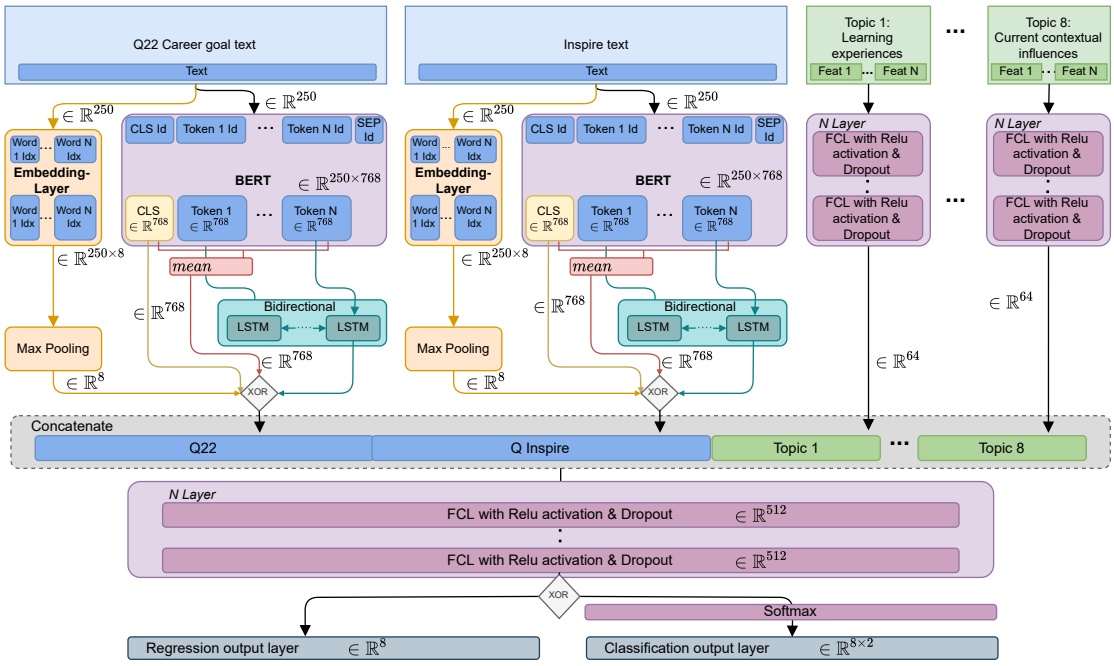


Figure 1: Model architecture combining both text and numerical (i.e. categorical) feature classification architectures. The XORs indicate different model choices for various sub-components.

tailed descriptions of these questions can be found in Gilmartin et al. (2017) or in a more condensed form in Appendix A of this paper.

While most of the questions in the survey are multiple-choice, referred to as *numerical* or *categorical*, two questions require open-text answers. *Q22* asks about the short-term plans of students within five years of graduating while the *Inspire* question, asks how the survey itself influenced the thought process of the students towards their career goals.

The independent variable we are trying to predict is *Q20* also named *Career goal* in the survey and asks for the likelihood of a person to pursue a career in 8 distinct circumstances, ranging from corporate employee to non-profit founder. Each of these cases is given a Likert score from 0 to 4 representing the likelihood from *highly unlikely* to *very likely*. In our model, we use both the numerical responses from the 8 topics as well as the free-text answers to predict career preferences.

3.2 Model Architecture

The architecture for the prediction task is illustrated in Figure 1 and can be split into three logical parts. The first section (top left) deals with the open text variables and is based on DistilBERT and embedding layers. The second input section (top right), processes the numerical features pertinent to each topic through a series of *Fully Connected* (FC) layers.

After being processed in parallel, the latent representations of each open-text question and each topic are concatenated and processed through another FC block, before generating the final prediction.

The output is generated by two distinct heads: a regression task trained on mean absolute error loss approximating the numerical values of the subquestions of *Q20* and a classification output trained with a cross-entropy loss, predicting general favorable or unfavorable tendencies. In each case, there are eight individual outputs for each prediction, one for each task.

Open-end text variables: The main part of the text processing architecture is based on DistilBERT (Sanh et al., 2019), which is utilized without fine-tuning to create text representations for the following layers. The four branching architecture choices in this part include **(1)** the use of the embedding vector encoding the CLS token, **(2)** mean averaging over word token embedding vectors (Wolf et al., 2020), **(3)** feeding the word token vectors through a BiLSTM layer (Graves and Schmidhuber, 2005) and **(4)** a single eight-dimensional embedding layer trained on the free-text task data.

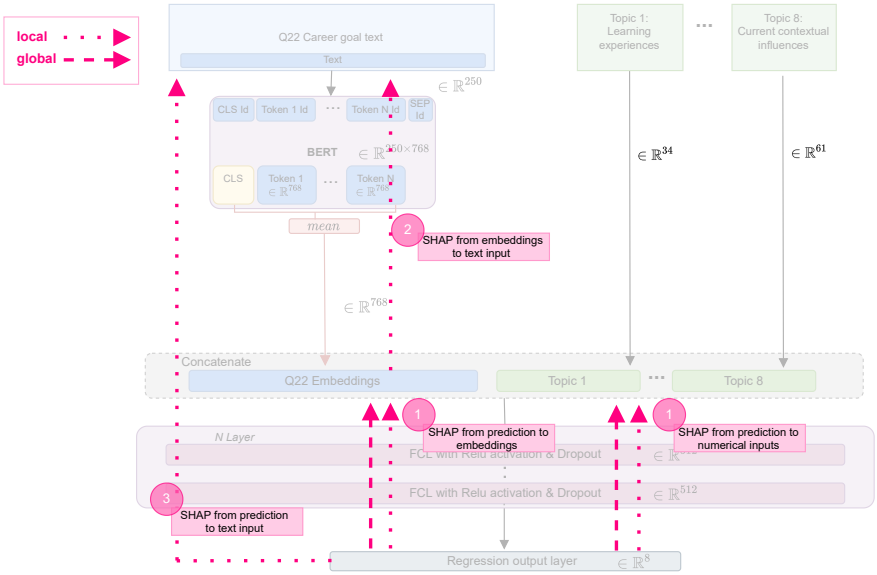


Figure 2: Explainability experiments with SHAP values for different parts of the model. **(1)** Global and local SHAP values from prediction to intermediate layer with embeddings and numerical features as inputs, **(2)** local SHAP values from embeddings to text input, **(3)** local SHAP values from prediction to text input

Numerical feature variables: This part of the architecture takes all recorded numerical features (minus the covariate) as input and groups them by topic according to the SCCT framework. Each topic is fed through separate FC layer model streams before being concatenated with the representation from the text variables. While most features can be input directly as a single value, some represent nominal choices and are input as one-hot encoding vectors instead.

3.3 Model Explanations

We apply several post-hoc explainability methods to both explain specific model predictions and gain a holistic understanding of what our model has learned.

Low-level feature and neuron explanations We employ SHAP (Lundberg and Lee, 2017) to compute local and global feature relevance explanations. This enables us to quantify the most important input components in terms of overall model accuracy, but also to identify the features dominating a specific prediction (Wich et al., 2021). Specifically, we **(1)** calculate and compare SHAP values for both the text and numerical value embeddings. Then, we **(2)** look at which parts of the text input trigger the neurons presenting the highest activation in the previous analysis. Finally, we **(3)** compute SHAP values for the input text w.r.t. the final model prediction. Figure 2 shows a detailed overview of all SHAP explanation experiments and how they relate to the various model inputs and inner components.

High-level concept explanations: We utilize ConceptSHAP (Yeh et al., 2020) to understand how the model captures and organizes higher-level information for its predictions. This information is extracted in the form of concepts, i.e. clusters of embedding vectors each summarized by a concept vector c_i which acts as the cluster's centroid. Beyond their extraction, we **(1)** use the K nearest neighbors of c_i to describe each concept, **(2)** measure the influence of each concept for a single prediction, and **(3)** report *completeness scores* - i.e. how well the set of extracted concepts describe the model's behavior (Yeh et al., 2020). Analogous to Figure 2 for SHAP experiments, Figure 12 (See Appendix C) shows a detailed overview of all ConceptSHAP explanation experiments and how they relate to the various model inputs and inner components.

4 Results

Results are presented in two distinct sections. Firstly, we present the numerical results for the prediction task in the case of both the regression and the classification heads for the whole architecture. The performance here is evaluated through

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

Table 1: F1 Scores for the combined model, utilizing different parts of the input data. Architectures differ based on which parts of the input they use. Question 22 (Q22) and Question Inspire (Ins.) are free text questions, tabular data (T) is counted separate. All numbers are reported for performance on classification (C) and regression (R) tasks. Best model for each task (T1 to T8) in bold.

macro F1 score for all eight individual topic predictions. Secondly, we show explanations for these model predictions through explainability frameworks SHAP and ConceptSHAP.

4.1 Task Performance

We conducted a variety of experiments on different sub-parts of the architecture and finally on different overall combinations of features for the architecture presented in Figure 1.

Text-based prediction We tested four different configurations of the free-text part of the model architecture, each with a different mode to generate embeddings as described in section 3.2. Results are taken individually for each of the eight tasks and for both regression and classification heads. A stripped-down version of these results for task 8 *Founding for-profit* can be found in Table 2. The full table of results can be found in Appendix D.

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

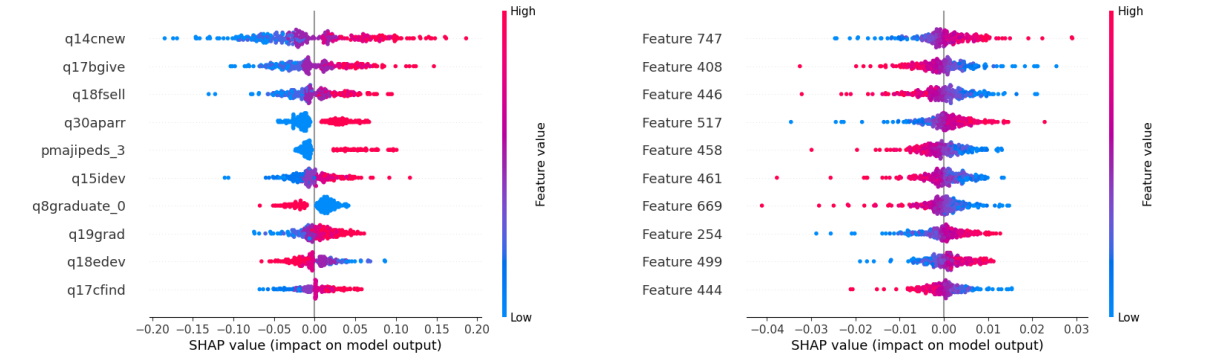
Table 2: F1 Scores for the Q22 text input, predicting task 8 (T8) for each architecture. Best model in bold.

In summary, the mean average model performed best on the label 8 task, scoring an F1 score of 63.70% for the classification and 62.40% for the regression task. On six of the other tasks, the *mean*-model performed better than the other models. The classification task was overall easier to achieve, yielding higher scores across all tasks with the notable exception of task 4.

Numerical variable-based prediction In this part of the evaluation, we ran the numerical variable part of the architecture without any text inputs to compare results on the 8 tasks (T1 to T8). We evaluated the input of each of the 8 SCCT topics individually, as well as on the combination of all topics for prediction.

The best performing model utilized all available topics concatenated directly before processing with a mean F1 score of 72.65% (C) for the classification and 74.65% (R) for the regression head on task 8. The full list of results is available in Appendix D. Based on the numerical variables only, it is unclear whether the classification or the regression head performed better overall since performance turned out to be highly task and architecture-dependent.

Combined performance The overall performance of the model is evaluated for a variety of feature combinations. For all the cases we chose the best performing combinations of the architecture for text-based prediction and the concatenated input of all SCCT topics for the numerical variable input. The combination of possible features is then for text input either *no text*, *Q22*, the *Inspire* ques-



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

(b) Global expl., text embeddings only

(c) Local explanation: all features

(d) Local explanation: text embeddings only

Figure 3: SHAP values for all features (left) and text embedding only (right). Global explanations (top) and local explanations (bottom). The higher in magnitude the value is, the more important a feature is for the model, while a positive value contributes to a prediction value of 1 and a negative value to a class value of 0. See appendix E for a larger scale version of (c) and (d).

tion, as well as all numerical topic variables or none of them resulting in 8 total possible combinations.

The full evaluation of these input variations is shown in Table 1. Best results are achieved by the model combining *Q22* text input with the full set of SCCT topics, resulting in a macro F1 score of 73.64% (C) for classification and 76.23% (R) for regression. The *Inspire* text variable instead contributes negatively across tasks as well as scoring the worst for singular performance at 42.69% (C) and 35.48% (R) F1 score. Our best model thus uses all available numerical features, as well as the free-text input from *Q22* as input, processing the DistilBERT embedding into a mean sentence embedding vector and a regression head output for prediction.

4.2 Interpretability examples

For simplicity, we present explanations for the model reporting the best performance (see Table 1). For the first set of feature attribution explanations, we focus on the eighth head—capturing the *likelihood of starting a for-profit company*. For the concept-based explanations, instead, we examine all heads as concepts describe the information captured by the model overall.

Low-level feature and neuron explanations We begin by looking at the global importance of numerical features and text embeddings w.r.t. the model prediction. As one can see in Figure 3, the ten most important features are numerical features and no single embedded word is as relevant for the model. This is coherent with the observation in section 4.1 that additionally considering text led only to a slight performance improvement. Moreover, we can observe that the four most relevant features are *q14new*, *q17give*, *q18sell*, and *q30aparr*, which are particularly related with entrepreneurial behavior.

Figure 3 also shows two local explanations resulting from the first experiment. These again show the SHAP values for the text embeddings and the numerical features. The colors indicate whether the features push the prediction in a positive (pink for class 1) or negative (blue for class 0) direction. The strength of each feature's contribution is indicated by the length of its corresponding segment. Taking variable *q14cnew* as an example, low feature values impact the model negatively, while high values impact it positively, while in-between feature values land in between those values.

Examples of local explanations generated by the second and third experiments are visualized in Figure 4. In particular, we can observe the text features' influence both on the most influential neuron identified in the first experiments (4a) and on the

Concept	Nearest neighbors	Word cloud
1	want to be successful. find a job my own business no thanks work hard ill do whatever. no concrete plans yet run my own business. no comments no idea	software (5), my (6), no (17), thanks (6), idea (5), company (5), have (6), work (7)
2	i want to attend medical school i plan to find a mechanical i am planning to be a product i plan on working as a i would like to go into manufacturing and continue education with goal i would first like to pursue doctoral degree having my own company i will be starting a career as an seeking law degree, to move into	I (63), my (13), work (10), plan (24), find (5), graduate (8), will (17), be (17), go (7), am (5), career (6), get (6), job (7), would (13), like (14), engineering (7), working (13)
3	business learn skills, turn hobbies into i hope to run my own business start a company overseas earn experience in a small .. either go into industry or go gain experience in the industry. would like to get into management own company when i have the expertise my feet in a start up company early a good paying job at a company that	company (19), my (13), industry (14), work (22), engineering (18), start (12), I (21), business (6), go (12), own (6), job (9), pursue (5), will (8), plan (6), engineer (5), get (7), degree (6), masters (5), working (13), be (5)
4	school within the next two years. work there for 3 years in the next five years i hope work abroad at some point. 5 to 6 years. at least the next two years, i there for at least three years. tentative at that point in time i want in the next five years i field at least once.	at (19), my (13), go (12), industry (14), work (22), engineering (18), start (12), I (21), business (6), engineer (5), be (5), own (6), job (9), pursue (5), will (8), plan (6), get (7), degree (6), masters (5), working (13)

Table 3: The four concepts with 10 examples from the top 100 nearest neighbors and the word clouds containing the most frequent words from the nearest neighbors

model's output (4b). It is instructive to notice that—in contrast to the model as a whole—SHAP values w.r.t. to this specific neuron are all non-negative. This indicates that this unit has specialized in capturing only positive features, i.e. desire to start a for-profit company.

Higher-level concept explanations While ConceptSHAP (Yeh et al., 2020) does not require a predefined list of concepts, we still need to manually set how many we want to model. We choose four as we are seeking to extract broad and general concepts.

For each concept, we look at the 100 nearest neighbors' word embeddings. We then map these back to their corresponding word token and include four neighboring tokens from their corresponding sentence. Furthermore, we count the word tokens appearing in the top 100 nearest neighbors and construct a word cloud with the ones occurring more than five times.

Once the concepts have been extracted automatically, they can be inspected manually by humans who can look for a common theme in the word cloud and the nearest neighbors. Table 3 presents an overview of the extracted concepts via showing the ten nearest neighbors in addition to the word cloud extracted from the top 100.

The first concept mainly contains nearest neighbors describing a lack of orientation and concrete career plans. Indeed, "no" is one of the words dominating this word cloud. The second, in contrast, captures a strong sense of having a clear path for the own future career. Here, most sentences start

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

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

Figure 4: Local SHAP values describing the impact of the embedding layer and numerical feature inputs on the model's prediction for 4 different samples, 2 belonging to class 0 (not wanting to start a for-profit company) and 2 belonging to class 1 (wanting to start a for-profitcompany). See E for a larger scale version.

with "I" and contain words like "will" and "plan", indicating strong traits of self-centeredness and determination. Both these concepts match what also discovered by Grau et al. (2016, p.8): i.e. the *clarity of plans*.

The third concept revolves around the plan type rather than its certainty or concreteness. For instance, we find general words like "company", "work", and "engineering", which indicate the goal of founding a company, joining a startup, or working in the industry. This matches the idea of *career characteristics*, also found in Grau et al. (2016, p.8). Finally, the last concept is the most distinctive as it captures the *plan timeline*, clearly present in all the nearest neighbors listed. This concept, connecting career plans to the time dimension, cannot be found in previous works such as Grau et al. (2016). The completeness scores achieved by these concepts are reported in the appendix (see C).

5 Discussion and Comparison

We employed several architectures to solve the the problem of career choice prediction to improve over prevailing closed and open-vocabulary methods. While for some survey responses correlations were strenuous, we found general success in predicting variables relating to entrepreneurial aspirations.

We see an overall increase in performance by combining textual and numerical input data. While numerical data is generally more predictive in our experiments, the 119 numerical variables are also a lot more nuanced than the free-text answers *Q22* and *Inspire*. Despite this, prediction from text alone still manages to perform relatively well across different tasks. The negative impact on performance of including the *Inspire* variable in models is likely due to the limited amount of text in the answers to the question.

To back up our model findings with explanations, we applied SHAP and ConceptSHAP as post-hoc approaches. The first confirmed what we observed in terms of model performance and provided us with a good understanding of the global and local relevance of each component: numerical features, text features, and embeddings. The second, instead, led to the identification of relevant concepts —*clarity of plans*, *career characteristics*, and *plan timeline*—in line with the human judgment of previous works.

6 Conclusion and Future Work

This work investigated the usage of state-of-the-art NLP and XAI techniques for analyzing user-generated survey data. Instead of manually examining individual answers, our methodology heavily relies on analyzing and interpreting a predictor model trained to extract correlations and patterns from the whole data set. We proposed a multi-modal architecture consisting of a DistilBERT transformer architecture and FC layers. The former is used to extract information from open-ended textual answers while the latter process the numerical features representing closed-ended answers. The model achieves satisfactory accuracy in predicting students' career goals and aspirations.

We leveraged SHAP and ConceptSHAP to generate both instance-level and concept-level explanations. These methods were applied at different levels of granularity to assemble a holistic understanding of the model's reasoning. Experiments on the EMS survey show promising results in predicting the students' entrepreneurial ambition. Moreover, local explanations provide us insights about the most relevant questions overall as well as relevant factors w.r.t. a single student. The automatic high-level concept analysis also led to insightful findings which were very similar to what was found in previous research including human judgment.

We release our code to the public to facilitate further research and development [1].

Acknowledgments

This paper is based on a joint work in the context of Katharina Hermann's master's thesis (Hermann, 2022).

[1] https://github.com/EdoardoMosca/explainable-ML-survey-analysis

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Sara A Atwood, Shannon K Gilmartin, Angela Harris, and Sheri Sheppard. 2020. Defining first-generation and low-income students in engineering: An exploration. In *ASEE Annual Conference proceedings*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Antony Bryant and Kathy Charmaz. 2007. *The Sage handbook of grounded theory*. Sage.

Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Johannes C Eichstaedt, Margaret L Kern, David B Yaden, HA Schwartz, Salvatore Giorgi, Gregory Park, Courtney A Hagan, Victoria A Tobolsky, Laura K Smith, Anneke Buffone, et al. 2021. Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4):398.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Shannon K Gilmartin, Helen L Chen, Mark F Schar, Qu Jin, George Toye, A Harris, Emily Cao, Emanuel Costache, Maximillian Reithmann, and Sheri D Sheppard. 2017. Designing a longitudinal study of engineering students' innovation and engineering interests and plans: The engineering majors survey project. ems 1.0 and 2.0 technical report. *Stanford University Designing Education Lab, Stanford, CA, Technical Report*.

Michelle Marie Grau, Sheri Sheppard, Shannon Katherine Gilmartin, and Beth Rieken. 2016. What do you want to do with your life? insights into how engineering students think about their future career plans. In *2016 ASEE Annual Conference & Exposition*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Timothy C Guetterman, Tammy Chang, Melissa DeJonckheere, Tanmay Basu, Elizabeth Scruggs, and VG Vinod Vydiswaran. 2018. Augmenting qualitative text analysis with natural language processing: methodological study. *Journal of medical Internet research*, 20(6):e9702.

Katharina Hermann. 2022. Explaining neural nlp models to understand students' career choices. Master's thesis, Technical University of Munich. Advised and supervised by Edoardo Mosca and Georg Groh.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 279–287.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Robert W Lent, Steven D Brown, and Gail Hackett. 1994. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior*, 45(1):79–122.

Amber Levine, T Bjorklund, Shannon Gilmartin, and Sheri Sheppard. 2017. A preliminary exploration of the role of surveys in student reflection and behavior. In *Proceedings of the American Society for Engineering Education Annual Conference, June 25-28. Columbus, OH.*

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *NeurIPS 2017*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102.

Scott Parrigon, Sang Eun Woo, Louis Tay, and Tong Wang. 2017. Caption-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *Journal of personality and social psychology*, 112(4):642.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep nlp models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *NeurIPS 2017*.

Mark Schar, S Gilmartin, Beth Rieken, S Brunhaver, H Chen, and Sheri Sheppard. 2017. The making of an innovative engineer: Academic and life experiences that shape engineering task and innovation self-efficacy. In *Proceedings of the American Society for Engineering Education Annual Conference, June 25-28. Columbus, OH.*

Maximilian Wich, Edoardo Mosca, Adrian Gorniak, Johannes Hingerl, and Georg Groh. 2021. Explainable abusive language classification leveraging user and network data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 481–496. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565.

A Appendix: Details on the EMS 1.0 survey data

The longitudinal *Engineering Major Survey* (EMS) by Gilmartin et al. (2017) consists of three surveys in total, conducted between 2015 and 2019. In this paper we only focus on the EMS 1.0 data from 2015 consisting of 7197 surveyed students of engineering enrolled at 27 universities in the US. The study is based on the *Social Cognitive Career Theory* (SCCT) framework (Lent et al., 1994) about how a students decision making is influenced by 8 specific topics.

These topics are:

- Topic 1: Learning experiences
- Topic 2: Self-efficacy (Engineering task, professional/interpersonal, innovation)
- Topic 3: Innovation outcome expectations
- Topic 4: Background characteristics / influences (gender, ethnicity, family background)
- Topic 5: Innovation interests
- Topic 6: Career Goals: Innovative work
- Topic 7: Job Targets
- Topic 8: Current contextual influences (major, institutional, peer)

Independent variables: Our independent variables come from topic 7 and surmise the following question *Q20*: "How likely is it that you will do each of the following in the first five years after you graduate?". It provides eight career possibilities which constitute our tasks 1 through 8 for each of the prediction heads:

1. Work as an employee for a small business or start-up company.

2. Work as an employee for a medium- or large-size business.

3. Work as an employee for a non-profit organization (excluding a school or college/university).

4. Work as an employee for the government, military, or public agency (excluding a school or college/university).

5. Work as a teacher or educational professional in a K-12 school.

6. Work as a faculty member or educational professional in a college or university.

7. Found or start your own for-profit organization.

8. Found or start your own non-profit organization.

Each entry can be answered with a Likert scale score ranging from 0 *'Definitely will not'* to 4 *'Definitely will'*.

For classification, the 5 classes (0 through 4) are binned into a binary label: low interest and high interest. The binning is done depending on the median of each label as illustrated in Figure 5. However this strategy ultimately still leads to unbalanced classes in some cases.

Lastly, we also analyze Pearson Correlation between all remaining labels after list-wise deletion, to determine whether they can be considered unique tasks. Our analysis illustrated in Figure 6 illustrated this point with most classes showing low correlation (less than 0.5).

Numerical variables: There are 119 numerical feature variables that operate on a categorical or five-point scale split across 30 distinct questions. Scale design, as well as the order of questions was based on minimizing bias in survey response.

An additional test of correlation between numerical features and task labels showed only weak linear correlation, indicating that solving the task is more complex.

Open text variables: We consider two open text variables, which are the following:

1. *Q22*: "We have asked a number of questions about your future plans. If you would like to elaborate on what you are planning to do, in the next five years or beyond, please do so here."

2. *Inspire*: "To what extent did this survey inspire you to think about your education in new or different ways? Please describe."

While these questions nominally fall under topic 7 in the SCCT framework, we treat them as disjoint topics during processing.

We additionally evaluated text length and correlation between the description of tasks of our target variable and the contents of the free text fields. Text length does not correlate with our label classes as shown in Figure 7. At the same time we could detect some correlation through keyword matching with *Q22*, especially relating to a lower score. Meanwhile there is no strong correlation between keywords for the Inspire variable. Results of the correlation analysis can be found in Figure 8 and Figure 9.

B Appendix: Non-combined architectures

This appendix shows the schematics for both architectures which omit either the textual or numerical variable part which was used for the detailed experiments listed in Appendix D. The text-only architecture can be found in Figure 10 while the numerical-only model can be found in Figure 10.

C Appendix: Higher-Level ConceptSHAP Experiments

Figure 12 shows an overview of the experiments involving ConceptSHAP (Yeh et al., 2020). Completeness scores for the retrieved concepts are reported in Table 4.

D Appendix: Detailed experiment results

This section lists the full results for the text-only classification and regression tasks across topics in table 5 as well as the results for the numerical variable prediction in table 6.

E Further SHAP Examples

To improve their readability, we now present again the SHAP force plots already included in 4.2. We also present further examples not previously included.

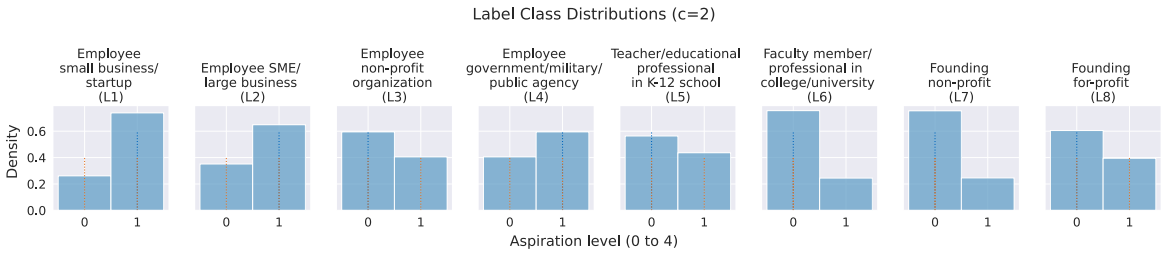


Figure 5: Splits binning 5 classes into two by median for each task.

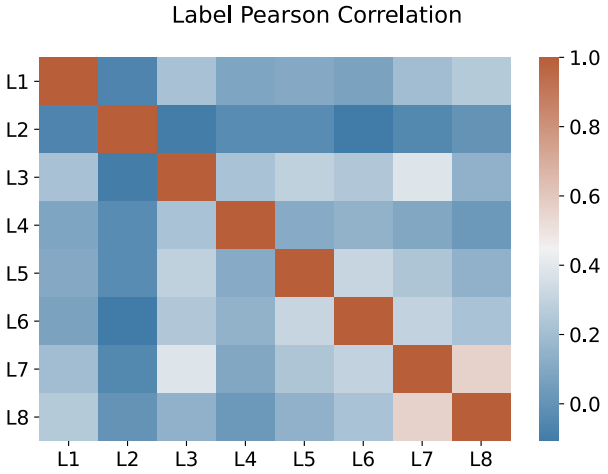


Figure 6: Pearson Correlation between each of the 8 labels. Values range from 0.0 to 1.0.

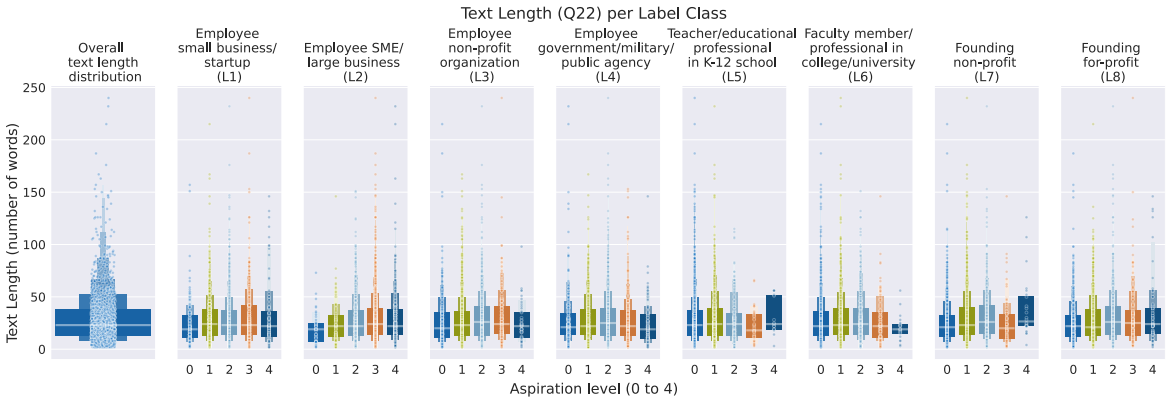


Figure 7: Overall text length distribution of Q22 and distribution grouped by classes per label.

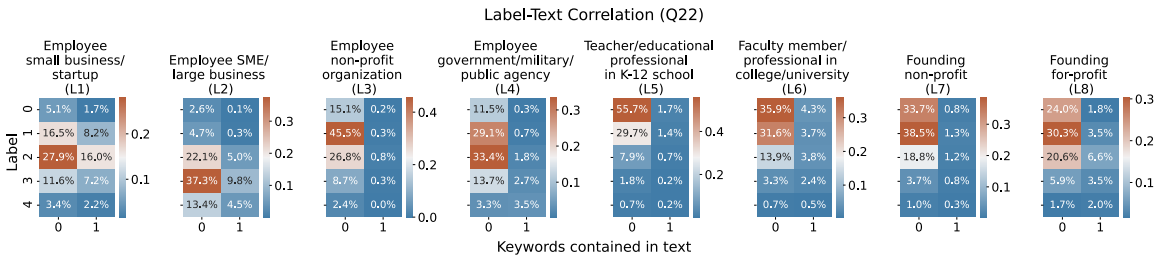


Figure 8: Model architecture for numerical features with FC layers.

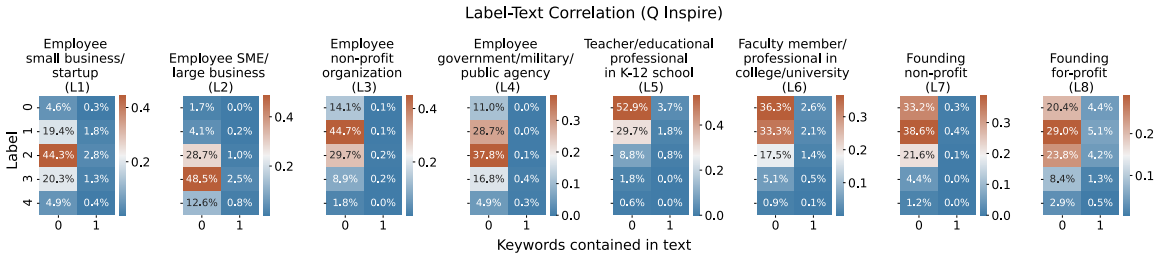


Figure 9: Model architecture for numerical features with FC layers.

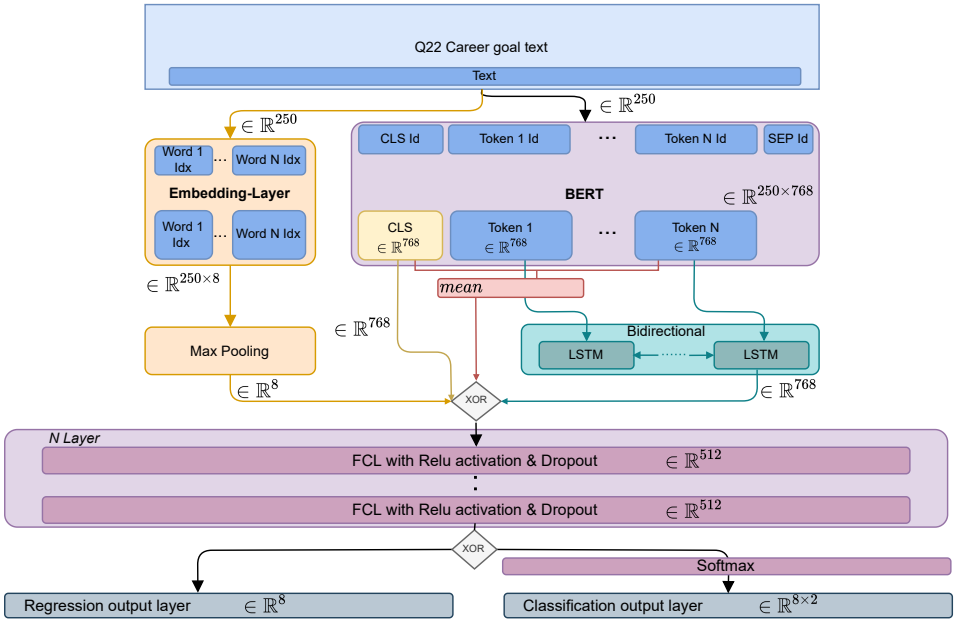


Figure 10: Model architecture for prediction through text processing. The XOR signifies different model choices w.r.t. different embedding processing steps and different output heads.

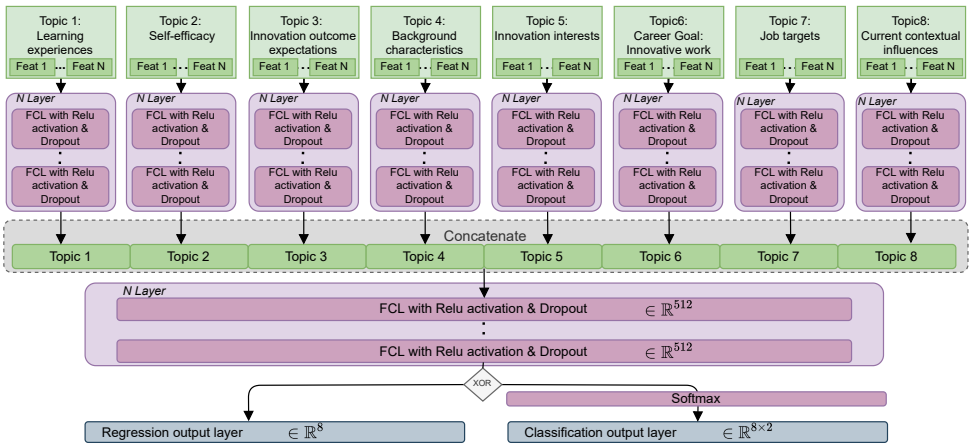


Figure 11: Model architecture for numerical features with FC layers. The XOR indicates the different model choices w.r.t. different output heads choices.

L1	L2	L3	L4	L5	L6	L7	L8
-0.66	-0.79	0.17	-0.59	0.18	0.93	0.89	0.73

Table 4: The completeness scores for each of the 8 prediction heads measuring how well the concepts can be used to recover predictions from the original model **(3)**

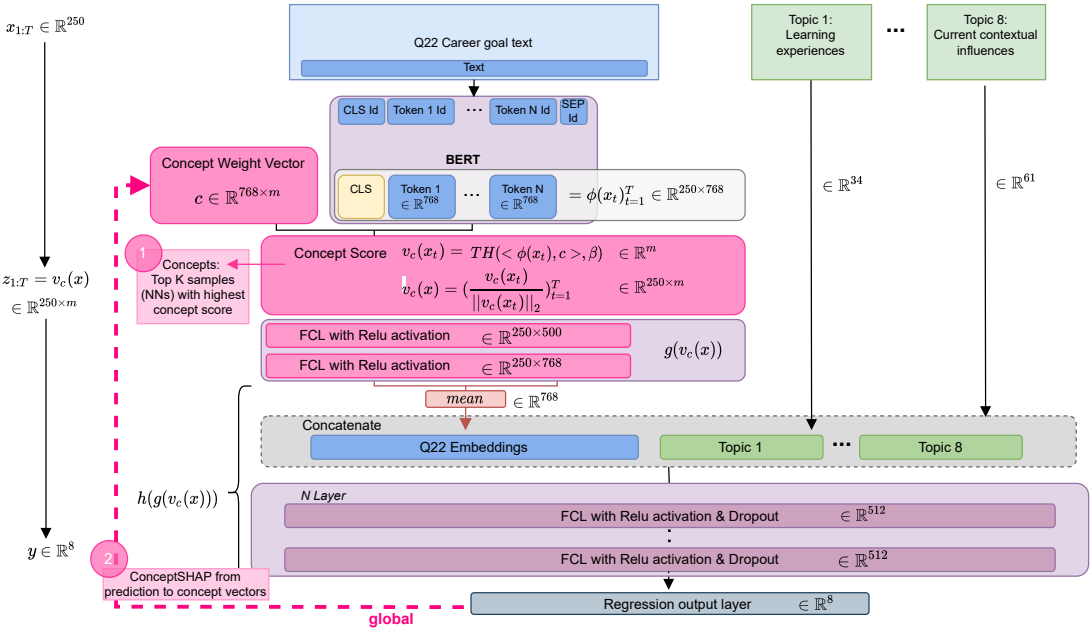
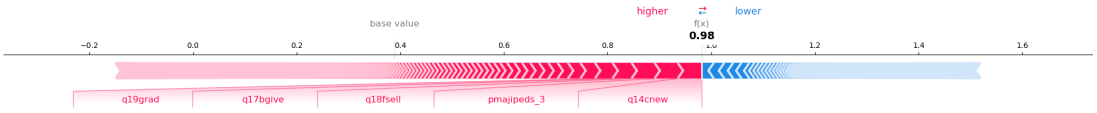


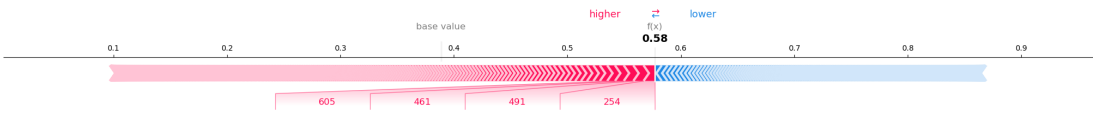
Figure 12: Explainability experiments with a concept-based method called ConceptSHAP. The original model is extended to a surrogate model to train concept vectors c_j , which function as the centroids of the concepts. These concepts are then being formed by the top k nearest neighbour tokens embeddings to the concept vectors (**1**). In addition to the pure concept extraction, we can measure their importance for the prediction of the model by using the principle of SHAP, (**2**).

		T1	T2	T3	T4	T5	T6	T7	T8
CLS	C	57.12	58.05	48.49	48.17	42.26	46.42	44.74	60.66
	R	54.05	51.26	36.41	44.24	35.21	42.74	43.44	53.96
mean	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
BiLSTM	C	42.75	38.74	39.17	37.73	35.36	43.11	42.18	37.88
	R	52.82	54.49	36.70	49.77	34.91	42.38	42.62	58.18
embedding	C	**54.57**	47.62	50.52	50.06	48.31	48.05	46.45	49.66
	R	52.21	47.68	47.83	43.04	48.06	43.56	51.22	50.27

Table 5: F1 Scores for the Q22 text input, predicting all tasks. Best model for each task in bold.



(a) Local explanation: all features



(b) Local explanation: text embeddings only

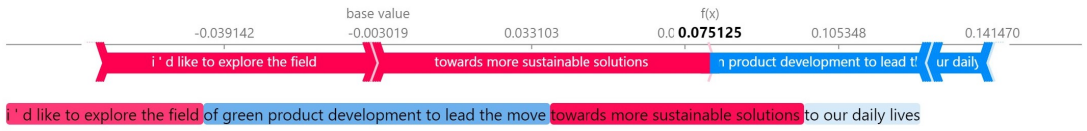
Figure 13: Larger scale version of plots (c) and (d) from Figure 3

			T1	T2	T3	T4	T5	T6	T7	T8
topic 1		C	44.39	41.74	57.46	41.36	52.21	49.62	58.07	66.83
		R	41.63	40.48	44.78	42.77	44.04	44.92	46.51	64.92
topic 2		C	48.42	44.83	54.36	42.51	40.32	42.60	42.74	62.39
		R	43.56	39.98	36.46	38.16	35.17	43.68	43.09	55.41
topic 3		C	42.74	46.80	48.03	42.17	54.85	46.05	48.10	50.18
		R	43.33	39.68	45.84	38.02	48.54	46.42	47.09	48.60
topic 4		C	42.28	39.33	45.18	47.16	45.22	44.94	44.71	54.37
		R	42.17	40.39	44.03	51.85	41.54	48.91	48.24	48.06
topic 5		C	44.94	51.00	58.68	45.98	55.64	46.77	43.44	64.33
		R	44.33	48.75	53.58	41.33	51.86	43.06	43.51	62.98
topic 6		C	49.26	40.35	44.68	47.18	38.39	42.71	42.57	57.70
		R	44.36	44.39	37.53	38.89	35.85	42.46	43.16	61.96
topic 7		C	46.40	61.60	56.66	46.20	54.11	58.03	43.02	44.29
		R	47.32	**62.69**	50.31	51.28	52.65	60.86	43.59	48.98
topic 8		C	46.41	44.39	52.06	51.84	45.58	45.68	44.04	48.69
		R	48.92	56.72	49.96	53.97	49.65	53.29	43.37	38.91
all topics	sep.	C	51.41	60.80	60.90	**57.35**	**61.06**	60.79	59.29	70.25
		R	**51.81**	55.66	52.38	56.31	52.84	55.83	53.32	67.74
	dir.	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	56.92	59.08	**74.65**

Table 6: F1 Scores for the numerical data differing on inputs only. Best model for each task in bold.

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

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



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

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

Figure 14: Larger scale version of SHAP plots presented in Figure 4. Two additional examples have also been added - i.e. (c) and (d).