Response Construct Tagging: NLP-Aided Assessment for Engineering Education

Ananya Ganesh[♢] **Hugh Scribner**[♢] **Jasdeep Singh**[♢]
Katherine Goodman[♠] **Jean Hertzberg**[♢] **Katharina Kann**[♢]
[♢]University of Colorado Boulder [♠]University of Colorado Denver
{ananya.ganesh, katharina.kann}@colorado.edu

Abstract

Recent advances in natural language processing (NLP) have greatly helped educational applications, for both teachers and students. In higher education, there is great potential to use NLP tools for advancing pedagogical research. In this paper, we focus on how NLP can help understand student experiences in engineering, thus facilitating engineering educators to carry out large scale analysis that is helpful for redesigning the curriculum. Here, we introduce a new task we call *response construct tagging* (RCT), in which student responses to tailored survey questions are automatically tagged for six constructs measuring transformative experiences and engineering identity of students. We experiment with state-of-the-art classification models for this task and investigate the effects of different sources of additional information. Our best model achieves an F1 score of 48. We further investigate multi-task training on the related task of sentiment classification, which improves our model's performance to 55 F1. Finally, we provide a detailed qualitative analysis of model performance.

1 Introduction

Engineering Education Research (EER) investigates effective pedagogical practices in engineering through qualitative and quantitative methods. A major focus of this research is curriculum design, particularly, inculcating "engineering thinking" (Moore et al., 2014; Pugh, 2002) and identity (Stevens et al., 2008) along with technical skills. In order to develop and improve such curricula, one effective method is to evaluate student experiences in engineering courses in a subjective manner, assessing several attributes such as their perception towards engineering in daily life, and the impact of the curriculum on their self-identity as an engineer (Clifford and Montgomery, 2015).

A popular framework to carry out such assessments is to administer surveys before and after completing a course, where students provide responses to carefully designed questions that probe for identity or affect (Sheppard et al., 2010). Surveys typically include some open-ended questions, such as *"How relevant is design for your intended career?"* to which students provide *text responses*. These are then manually analyzed to see, for example, whether students experience affective gain towards engineering after taking the course. In this paper, we propose using natural language processing (NLP) to enable educators to carry out this analysis faster and at a larger scale by automatically tagging student responses for their affective state towards pre-defined constructs which are of interest to educators.

We focus specifically on an industrial design course introduced in the mechanical engineering department of a large public university. Entry and exit surveys measure whether students undergo a *transformative experience* (Pugh, 2002) in the course, and assess the impact of the course on their

Construct	Description
Transformative Experience	
Expansion of Perception	The student sees everyday objects through the lens of course content
Motivated Use	The student applies ideas from course to everyday experiences
Affective Value	The student values course content for enriching everyday life
Engineering Identity	
Disciplinary Knowledge	The student displays grasp of technical concepts
Identification	The student sees themselves as an engineer
Navigation	The student sees their path towards becoming an engineer

Table 1: Descriptions of the constructs towards which affective state is classified.

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 250 - 261

engineering identity (Stevens et al., 2008). These aspects are characterized by six specific constructs, listed in Table 1. We introduce a new task, **response construct tagging (RCT)**, in which the goal is to identify student affect towards all six constructs from an open-ended response. For example, if a student response says *"I'm not sure what specific career I will pursue, but as long as it's engineering, I'm fine with it."*, then, they are displaying a positive affect towards the *Identification* construct since they see themselves as an engineer. Table 2 contains more examples of responses and human-annotated affect labels towards specific constructs.

Concretely, for each response, the RCT task is to classify affect corresponding to each of the six listed constructs. Our data consists of 232 student responses, annotated by a trained human annotator.[1] We investigate how NLP can be used to solve RCT, focusing on three research questions: 1) What is the most suitable NLP model for RCT? 2) What information relevant to the survey needs to be encoded? 3) Can other NLP tasks – specifically, sentiment classification – help with RCT through multi-task learning? We experiment with a classification model based on RoBERTa (Liu et al., 2019), a state-of-the-art language representation model, which achieves a score of 48 F1, and outperforms several baselines. We also find that multitask learning (Caruana, 1993) is highly effective, helping the classifier achieve an improvement of 6 points, from 48 F1 to 55 F1. Finally, we provide a detailed qualitative analysis of our model, looking at performance on individual survey questions, as well as errors made by the model.

2 RCT: Background and Task Description

2.1 Assessment in EER

Engineering Education Research (EER) is a field of inquiry (Jesiek et al., 2009; Froyd and Lohmann, 2014) that investigates and improves pedagogical practices in engineering disciplines, with the goals of increasing learning and student retention, including that of underrepresented groups (Prados, 1998). Research methodology in EER includes quantitative, qualitative and mixed-methods research (Borrego et al., 2009). Quantitative methods use statistics to study relationships between variables (such as class sizes) and outcomes (such as GPA). Qualitative research complements the above through analysis of data such as surveys and student interviews, which are frequently textual.

Several works discuss the value of qualitative studies for assessing educational practices (Borrego et al., 2009; Koro-Ljungberg and Douglas, 2008). Particularly, Olds et al. (2005) discuss the role of surveys, in which subjects self-report their experiences through open-ended or selected responses. Responses on surveys can be used to assess the effectiveness of various aspects of the engineering curriculum (Froyd et al., 2012), such as students' engagement. Educators are also interested in assessing whether the curriculum changes student perceptions of engineering as applied to their daily lives (Goodman, 2015), also known as undergoing a *transformative experience*. Another aspect of interest is the effect of the curriculum on the *engineering identity* of a student, i.e., whether the student sees themselves "becoming an engineer" (Stevens et al., 2008) in addition to acquiring technical skills. Entry and exit surveys before and after undertaking a course can indicate if the course resulted in affective gain towards such aspects. By analyzing student responses, educators can then redesign engineering curricula to promote such learning experiences, thereby increasing student motivation and retention (Baillie and Fitzgerald, 2000).

2.2 Industrial Design Course Survey

In this work, we look at an industrial design course at a large public university, which encourages students to use their engineering skills to create aesthetics-based design (Goodman et al., 2015). To assess the effect of the class on students, the instructors administer a 68-item survey (Sheppard et al., 2010) to students at the beginning and end of the course. Here, we describe only the open-response questions and the corresponding constructs they measure Example responses for each question, along with some of the constructs and corresponding affect, can be found in Table 2.

Open-response questions. The survey contains four open-ended questions, designed to elicit responses through which the specified constructs can be measured: **Q1)** *What motivates you when choosing an aesthetic while designing something?* This question helps us understand how students perceive the importance of design over pure functionality.

[1] Data, code and models can be found here https://nala-cub.github.io/resources/

Question	Response	Construct	Affect
What motivates you when choosing an aesthetic while designing something?	How will someone interact/ feel with this product. What emotion will it evoke.	Expansion of Perception	Positive
	I mostly focus on what will be the most functional aesthetic.	Motivated Use	Negative
How does making things on your own make you feel at the beginning of the process? Why does it make you feel that way?	I love the beginning of making things. Brainstorming and concept generation are some the most fun I have had in engineering.	Identification	Positive
	It makes me feel a little clueless, mostly because I always assume that there is a better or "perfect" way to carry out my design.	Disciplinary Knowledge	Negative
Are aesthetics important to the career you intend to pursue after graduation? Explain. Feel free to include what career you are interested in.	Very important, I am pursuing a career in human-centered design. My first job after college is as a Footwear Concept Engineer at Nike!	Navigation	Positive
	I am not sure what career I will be working in, but I know I enjoy design so aesthetics will be important to my career.	Affective Value	Positive
Are aesthetics important in your non-professional life? Explain.	No, I'm a pretty plain Jane. My walls are bare and I have no non-functional decorations.	Affective Value	Negative
	Personally, they aren't. However, I believe they would be if I had more disposable income.	Expansion of Perception	Neutral

Table 2: Examples from the industrial design course survey, with human-annotated affect labels.

Q2) *How does making things on your own make you feel at the beginning of the process? Why does it make you feel that way?* The purpose of this question is to gain insight into the ideation process. **Q3)** *Are aesthetics important to the career you intend to pursue after graduation? Explain. Feel free to include what career you are interested in.* Responses to this question shed light on whether students see themselves pursuing engineering careers. **Q4)** *Are aesthetics important in your non-professional life? Explain.* This question tells us whether students think of applying aesthetic design in their daily lives.

Constructs. We are interested in determining if students undergo a *transformative experience*, and whether the course has an impact on their *engineering identity*. Transformative experience can be characterized by three constructs: expansion of perception, motivated use and affective value.

- **Expansion of Perception:** the realization that how you view the world has changed due to the content you learned from the course. Students indicate this by observing learned concepts in their day-to-day lives.
- **Motivated Use:** the ability and desire to apply classroom learning to daily lives. Students indicate this by using ideas from courses without prompting in work or personal lives.
- **Affective Value:** the realization that learned concepts have some value in the real world. Students thus indicate a positive emotional state towards the course.

Engineering identity can be characterized by three constructs: disciplinary knowledge, identification, and navigation.

- **Disciplinary Knowledge:** the student indicates knowledge of concepts that engineers know. Additionally, the student thinks they can do what engineers do, and apply learning to the real world.
- **Identification:** the student indicates being identified as an engineer by themselves or others, which fosters a sense of belonging within the student towards engineering.
- **Navigation:** the student indicates their perception of how they are doing at becoming an engineer. This includes completing engineering-related coursework, and pursuing engineering internships or jobs.

Affect. Responses can indicate either `positive`, `negative`, or `neutral` affect towards a particular construct, as shown in Table 2. Responses that do not discuss a construct, or contain no affect information are annotated as `unavailable`.

2.3 Formal Definition of RCT

To automatically identify student affect towards constructs, we introduce the task of response construct tagging (RCT). We define this as a classification task, where, given a student response r together with a construct c, the goal is to predict the student's affect a towards c as expressed in r.

In this paper, $c \in$ {Expansion of Perception, Motivated Use, Affective Value, Disciplinary Knowledge, Identification, Navigation} and $a \in$ {Positive, Negative, Neutral, Unavailable}.

3 Datasets

3.1 Survey Data

Our data consists of 232 anonymized responses across all four questions from 29 students, both before and after completing the course. These responses were then annotated for affect by a trained human annotator for all six constructs.

We create training, development and test splits from 50%, 17%, and 33% of the data, containing, respectively, 114, 40 and 78 responses. Since each response is annotated for six constructs, we create six *training instances* from each response, where a training instance consists of the response and the construct name as input, and the affect label as the output. This finally gives us training, development and test sets of sizes 708, 240, and 468 respectively.

The distribution of labels in the training set is shown in Figure 1. We see that the labels are not evenly distributed – 480 responses, or 68% of the data, do not display any affective state, and are labeled as `unavailable`. Of the other labels, 174 responses, or 24%, are labeled as `positive`, 29 responses as `neutral`, and only 25 responses, or 3.5% of the data are labeled as `negative`. Further, Figure 1 also shows how the distribution of labels corresponds to the six constructs – we see that for several constructs, particularly those corresponding to *transformative experience*, no affect can be detected in the responses.

Table 3 shows the average statistics of responses in our training set, corresponding to the four affect labels. We see that responses annotated as `positive` are longer than others, containing more sentences and more tokens on average. We also compute the percentage of tokens that overlap between our responses and the Bing Liu sentiment lexicon (Hu and Liu, 2004), which contains word lists corresponding to positive and negative sentiment. Responses annotated as `positive` have a 5.82% overlap with the positive lexicon, however, `neutral` responses and those with no affect have more of an overlap, 6.25% and 6.22% respectively. With the responses that express `negative` affect, only 1.16% of tokens overlap with the negative lexicon. We compare this with a prototypical sentiment analysis dataset, also containing classroom survey responses (Welch and Mihalcea, 2016) in the last two rows of Table 3. Here, positive responses have a 9.67% overlap with the positive lexicon, while negative responses have a 4.59% overlap with the negative lexicon on average. This indicates that the affective states we are interested in are different from sentiment.

Feature	**Pos.**	**Neg.**	**Neu.**	**NA**
Num sentences	2.17	1.65	1.92	1.86
Num tokens	39.4	26.9	37.1	33.1
Pos. lexicon % overlap	5.82	4.90	6.25	6.22
Neg. lexicon % overlap	0.75	1.16	0.60	1.15
EECS				
Pos. lexicon % overlap	9.67	4.82	3.36	
Neg. lexicon % overlap	0.91	4.59	1.80	

Table 3: Average statistics of training set responses.

3.2 EECS data

Transfer learning via multitask learning (Caruana, 1993) has been shown to be successful in NLP (Collobert and Weston, 2008; Ruder, 2017). We therefore make use of the Michigan EECS Targeted Sentiment Analysis Dataset (Welch and Mihalcea, 2016) for training our model in an MTL setup. This dataset consists of student feedback from the Computer Engineering program posted on an online forum. Since responses may refer to either the course material or to the instructor, all responses include gold annotations for the entities mentioned in them. Responses are explicitly annotated for positive and negative sentiment, with the absence of annotations indicating neutral sentiment. The dataset contains a total of 1144 responses, from which we create training, development and test sets of sizes 645, 121, and 378 respectively.

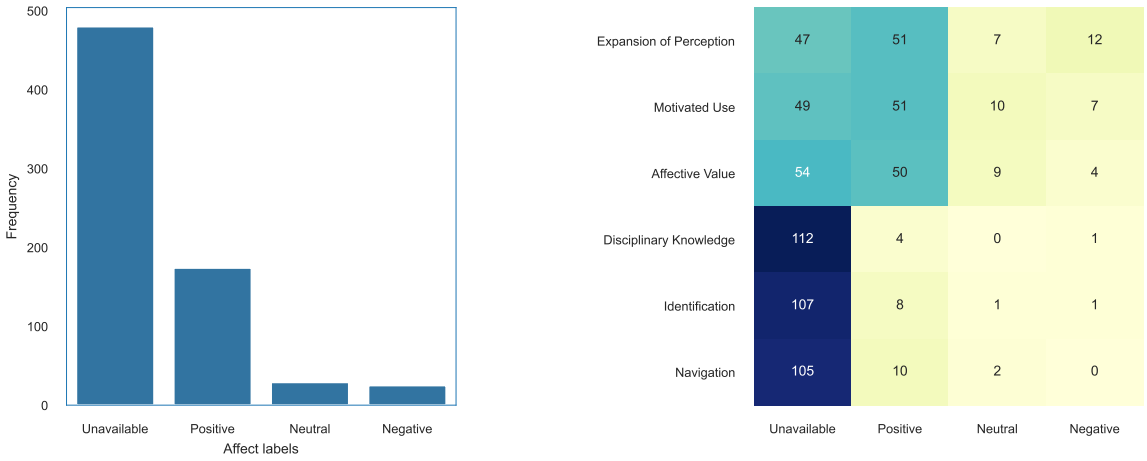

Figure 1: Distribution of affect labels in our training set.

4 Models

4.1 RoBERTa Classifier

Pretrained language representation models (Devlin et al., 2019; Liu et al., 2019) define the state-of-the-art on many language understanding tasks, including text classification (Wang et al., 2018). We thus finetune the RoBERTa model (Liu et al., 2019) for sequence classification, using the HuggingFace Transformers library (Wolf et al., 2020).

We train all models with a cross-entropy loss. We use the default hyperparameters of RoBERTa-base, with an embedding size of 512 and a hidden layer size of 768. We use a dropout probability of 0.1 on the attention layers and the hidden layers. We train for 50 epochs with early stopping on the development set, using the Adam optimizer (Kingma and Ba, 2014) and a learning rate of 1e-5. Training time was 10 minutes on a single nVidia V100 GPU.

4.2 Multitask Learning

Multitask learning (Caruana, 1993) enables models to learn from a similar task, and has been successfully used in NLP, particularly for tasks with a limited amount of data (Ruder, 2017; Benton et al., 2017; Mrini et al., 2021). We therefore perform multitask training on two tasks, namely RCT and sentiment classification on student course feedback. We use the Michigan EECS Targeted Sentiment Analysis Dataset (Welch and Mihalcea, 2016), as described in Section 3.2. This is done by jointly training a single model across both tasks, with a shared encoder and two separate classification heads.

5 Experiments

5.1 Baselines

Random The random baseline randomly selects one out of the four affect labels.

Majority The majority baseline predicts the label of the majority class, which is Unavailable.

Bag-of-Words + SVM Our final baseline represents each input response and construct as a bag-of-words. We vectorize the input using the Tf-idf vectorizer from scikit-learn (Buitinck et al., 2013). We then train an SVM classifier with a hinge loss, L2 regularization penalty of 1e-4, and a learning rate of 1e-5.

5.2 Additional input

We experiment with passing additional input available in our data – specifically, the question corresponding to a response, and the description of a construct as per the annotation guideline. As an example, for the construct *Navigation*, the description is "A response is tagged Positive for navigation if it discussed how the student felt that they were doing things that engineers do, such as accepting a position as a full-time engineer after graduation. Responses are marked as having negative navigation only if not feeling like an engineer was expressly mentioned". The complete list of descriptions can be found in the appendix. The additional input signals are concatenated to the text response before passing it to the model.

5.3 Metrics

For all models, we report accuracy, precision, recall, and F1 score. We compute F1 for all

Model	Acc.	Prec.	Recall	F1	Positive F1	Negative F1	Neutral F1	Unavailable F1
Random	23.7	24.4	21.8	18.2	23.5	11.7	2.7	34.8
Majority	68.1	17.0	25.0	20.2	0.0	0.0	0.0	81.1
BoW–SVM	67.5	48.7	36.2	35.7	45.5	5.8	10.5	81.0
RoBERTa	74.3	45.7	45.9	45.8	61.7	29.5	5.8	86.3
RoBERTa–Questions	77.9	49.8	47.4	48.4	64.9	33.3	6.9	88.7
RoBERTa–Question–Description	73.7	46.3	44.4	44.2	58.4	14.6	18.1	85.8
RoBERTa-Upsample	78.6	50.5	45.8	46.4	**66.9**	15.3	14.2	88.9
RoBERTa-Class-Weights	74.1	42.4	42.4	42.2	60.4	15.6	5.7	87.2
RoBERTa-MTL	79.2	65.8	54.2	**55.1**	62.0	**49.2**	**20.0**	**89.4**

Table 4: Model performance on the test set of RCT. Bold values indicate the model with the highest macro-averaged F1 on a specific category.

four classes individually, and additionally calculate macro-averaged F1 as an overall score for our dataset.

5.4 Balancing Classes

To counteract the label imbalance in our dataset, we experiment with two strategies: class weighting, and upsampling. With class weighting, we calculate weights for each output class, inversely proportional to its frequency in the training data, and use these weights while computing the cross-entropy loss. With upsampling, we repeat instances of the less frequent classes multiple times in the training set, such that all output classes are evenly represented.

5.5 Results

Table 4 shows the performance on our test set. Looking at the baselines, we see that while the *Random* and *Majority* baseline are comparable, the *BoW-SVM* baseline outperforms them by 15 F1. However, looking at the performance of the RoBERTa model, we can see that better input representations from pretraining makes a dramatic difference: RoBERTa outperforms our strongest baseline by 10 F1. Looking at performance on individual labels, we see that the model predicts with a high accuracy the labels which are dominant in the training set, i.e., `unavailable` and `positive`. However, it is negatively impacted by class imbalance – on the rarest label, `neutral`, it scores 5.8 F1, and on `negative`, it scores 29.5 F1.

Next, we incrementally encode additional input signals with both models as described in Section 5.2. We observe that encoding the question is effective, and overall performance increases to 48.4 F1 with RoBERTa. Particularly, on the `negative` affect label, performance increases by 4 F1. However, we find that additionally including the description of the construct doesn't result in further overall improvement over encoding the question.

We investigate two strategies for counteracting class imbalance as described in Section 5.4, namely upsampling and class weighting. We observe that upsampling has a positive effect on the rarest label, `neutral`, where RoBERTa performance goes up from 6.9 F1 to 14.2 F1. However, on the other labels, there is either a drop in performance or no noticeable change. On the other hand, class weighting does not result in improvement on any of the rarer classes, or overall.

Finally, we observe that the multi-task learning model achieves the highest performance on RCT, with an F1 of 55.1. Comparing to the equivalent single-task model RoBERTa-Questions, the MTL model improves by 6.7 F1, from 48.4 F1 to 55.1 F1. We also see a steep improvement on the rarer classes – on `negative`, performance improves by 15.9 F1, from 33.3 F1 to 49.2 F1, and on `neutral`, performance improves by 14.2 F1, from 5.8 F1 to 20.0 F1. Our results thus indicate that jointly training on the related task of sentiment classification helps the model learn the affect labels in our data better.

6 Analysis

6.1 Question-Level Performance

Figure 2 breaks down the performance of our best model, RoBERTa-MTL, across the four questions

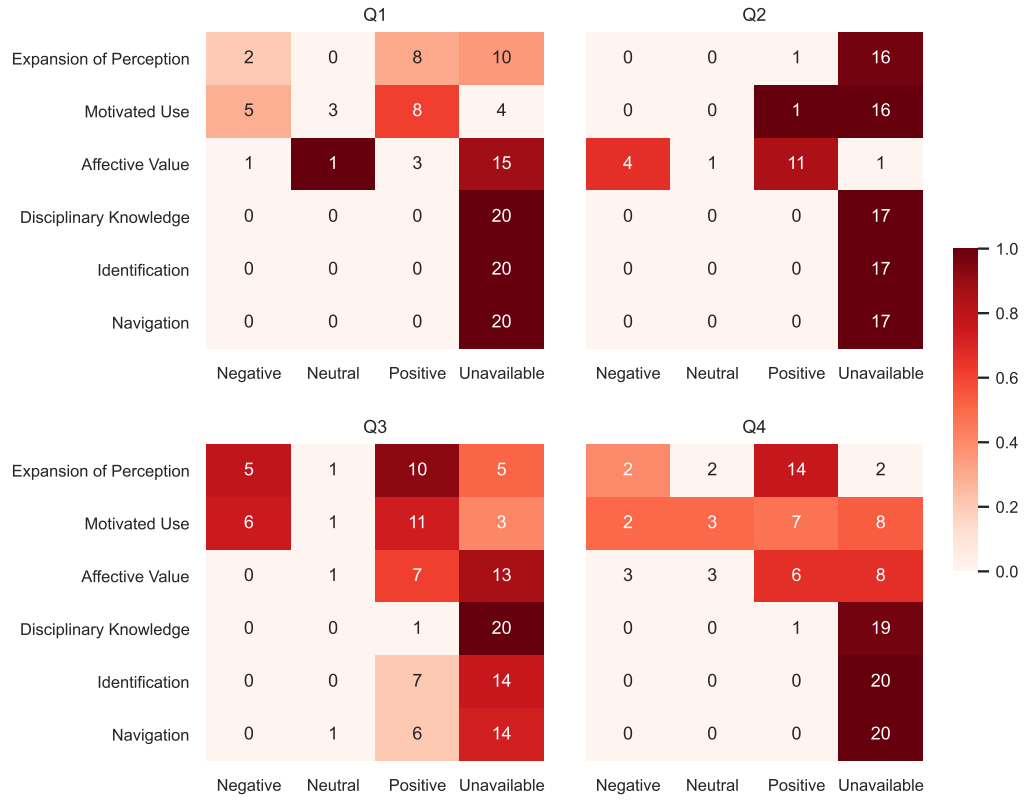

Figure 2: Performance of RoBERTa-MTL across questions for each construct and affect. Cell color intensity indicates F1, and cell values indicate label count.

on the survey as described in Section 2.2. The cell values show the count of each affect label corresponding to a construct on the test set, and the color intensity shows the model's F1 on a scale of 0–1. Overall, we see that the most frequent label is `unavailable`, on which the model's performance is also high across all constructs. This is particularly noticeable for the constructs of *Disciplinary Knowledge, Identification*, and *Navigation*, for which almost all annotations fall under `unavailable` except for Q3.

Next, we look at the plots for individual questions. We observe that for Q1, for the *Motivated Use* construct, the model does well on the `positive` and `negative` affect labels, but does not predict the other two classes correctly. For the *Affective Value* construct, the model predicts the `neutral` label correctly, but not `positive` or `negative`. For Q2, which asks students about their feelings towards starting a project, both `positive` and `negative` affect labels are frequent, and the model also performs well on these classes. On Q3, for both the constructs of *Expansion of Perception* and *Motivated Use*, F1 is high on both the `positive` and `negative` affect labels, but lower on the `unavailable` and `neutral` labels. Finally, for Q4, the model does well on the `positive` and `negative` labels for *Expansion of Perception*. On *Motivated Use*, the model achieves comparable performance on all labels, and on the *Affective Value* construct, the model does poorly at predicting `neutral` and `negative` labels.

6.2 Qualitative Analysis

We also look at specific errors made by the model on the test set, as shown by the confusion matrix in Figure 3. We observe that the major source of error is from confusing a true class of any label with the `positive` label. An example of a true `negative` response to Q1 being predicted as `positive` towards *Motivated Use* is: *"I generally just design to my own tastes and hope that its appealing to others"*. Here, the student indicates that they do not make use of learned concepts while designing which indicates negative affect towards *Motivated Use*. However, this could be perceived as positive, since the student indicates an interest in design, and potentially due to the use of the word "appealing", which typically

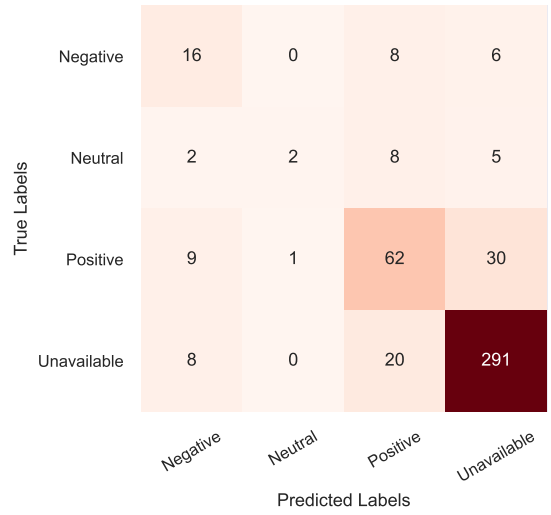

Figure 3: Confusion matrix of RoBERTa–MTL predictions.

co-occurs with positive text for sentiment classification. We also observe that when the true class is `positive`, the majority errors are due to predicting `unavailable` or `negative`. An example of a true `positive` response, predicted as `negative` is, in response to Q4: *"It is important sometimes, like when I'm trying to decorate my house or choosing an outfit to go out in."*

7 Related Work

Prior research has investigated how NLP can be used to analyze student feedback, with the goal of improving teaching and learning. Similar to our work is sentiment analysis for classifying student's affective states after completing a course (Dolianiti et al., 2018; Kastrati et al., 2021). More specifically, aspect-based sentiment analysis (Pontiki et al., 2016) is used to determine sentiment towards distinct entities such as instructors or course material (Ramesh et al., 2015; Welch and Mihalcea, 2016), as well as attributes such as teachers' helpfulness (Nikolić et al., 2020) or quality of examples used (Chathuranga et al., 2018). Several methods have been investigated for this problem, including sentiment lexicons (Welch and Mihalcea, 2016; Wen et al., 2014b), probabilistic models (Ramesh et al., 2015), convolutional neural networks (Kastrati et al., 2020), and LSTM models (Nguyen et al., 2018). However, our proposed task differs from aspect-based sentiment analysis since the constructs we are looking for are implicit, and are never explicitly mentioned in a student response.

Beyond sentiment classification, other applications have been studied for understanding student feedback: Luo and Litman (2015) automatically summarize student responses to open-ended reflection prompts, and Luo et al. (2016) summarize student feedback on courses. Wen et al. (2014a) analyze posts on MOOC forums to determine student motivation and engagement. In engineering education research, NLP has been used for determining "disciplinary discourse" in student resumés (Berdanier et al., 2018), and for measuring metacognitive development of students in engineering classrooms (Bhaduri, 2018).

In our experiments, we use pretrained models for classification through fine-tuning, which have proven to be highly effective for NLP tasks (Wang et al., 2018, 2019). Pretrained models have also been used successfully in educational applications (Alikaniotis and Raheja, 2019; Benedetto et al., 2021; Katinskaia and Yangarber, 2021). We also use multi-task learning (Caruana, 1993), which has been investigated for tasks such as text classification (Liu et al., 2017) and sequence labeling (Hu et al.; Bingel and Søgaard, 2017). Multi-task learning has proven to be particularly effective in low-resource settings (Benton et al., 2017; Schulz et al., 2018; Mrini et al., 2021), which is applicable for our task as well.

8 Conclusion

We introduce a new task, response construct tagging, to automatically tag student survey responses for the affective state of a student towards six predefined constructs. We present a classification model for this task based on the RoBERTa pretrained model, that outperforms multiple baselines. On investigating the different information sources this model can utilize, we find that the best performance of 48.4 F1 can be attained by encoding a response, construct, and the corresponding question. We also demonstrate the benefits of training our model in a multitask learning setting on the related task of sentiment classification, which achieves a score of 55.1 F1, a 6.7 F1 improvement. Our task, and corresponding model, enables educators to assess the effectiveness of their curriculum in influencing students' identity and perceptions of engineering, and thereby to design curricula that maximize positive influence.

Limitations and Future Work Our proposed model can detect certain constructs and affects with

high accuracy, such as the positive labels. However, RCT is a challenging task – differences between affect labels are nuanced, and a single response can indicate different affective states towards different constructs. Moreover, the sparsity of labels in our dataset makes it difficult to learn the rarer combinations of affect and constructs, such as negative *Identification*. However, this is an inherent limitation with the classroom assessment framework, since students might be unwilling or unlikely to express feelings such as "not identifying as an engineer". One way to mitigate this problem might be to generate student responses artificially for constructs and affects that are not represented in the dataset. In future work, we will investigate how this can be done both manually, i.e., using human annotators, and automatically, such as conditionally generating responses that display a desired affect.

Acknowledgments

We would like to thank the reviewers for their feedback and suggestions. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF. This research was also supported by the Interdisciplinary Research Themes initiative at the University of Colorado Boulder.

References

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–133, Florence, Italy. Association for Computational Linguistics.

Caroline Baillie and Geraldine Fitzgerald. 2000. Motivation and attrition in engineering students. *European Journal of Engineering Education*, 25(2):145–155.

Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–157, Online. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Catherine G.P. Berdanier, Eric Baker, Weiqin Wang, and Christopher McComb. 2018. Opportunities for natural language processing in qualitative engineering education research: Two examples. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–6.

Sreyoshi Bhaduri. 2018. *NLP in Engineering Education-Demonstrating the use of Natural Language Processing Techniques for Use in Engineering Education Classrooms and Research*. Ph.D. thesis, Virginia Tech.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Maura Borrego, Elliot P Douglas, and Catherine T Amelink. 2009. Quantitative, qualitative, and mixed research methods in engineering education. *Journal of Engineering education*, 98(1):53–66.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Janaka Chathuranga, Shanika Ediriweera, Ravindu Hasantha, Pranidhith Munasinghe, and Surangika Ranathunga. 2018. Annotating opinions and opinion targets in student course feedback. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Valerie Clifford and Catherine Montgomery. 2015. Transformative learning through internationalization of the curriculum in higher education. *Journal of Transformative Education*, 13(1):46–64.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Foteini S Dolianiti, Dimitrios Iakovakis, Sofia B Dias, Sofia Hadjileontiadou, José A Diniz, and Leontios Hadjileontiadis. 2018. Sentiment analysis techniques and applications in education: A survey. In *International Conference on Technology and Innovation in Learning, Teaching and Education*, pages 412–427. Springer.

Jeffrey E. Froyd and Jack R. Lohmann. 2014. *Chronological and Ontological Development of Engineering Education as a Field of Scientific Inquiry*, page 3–26. Cambridge University Press.

Jeffrey E Froyd, Phillip C Wankat, and Karl A Smith. 2012. Five major shifts in 100 years of engineering education. *Proceedings of the IEEE*, 100(Special Centennial Issue):1344–1360.

Katherine Goodman, Hunter Porterfield Ewen, Jiffer W Harriman Jr, and Jean Hertzberg. 2015. Aesthetics of design: A case study of a course. In *2015 ASEE Annual Conference & Exposition*, 10.18260/p.23504, Seattle, Washington.

Katherine Ann Goodman. 2015. *The transformative experience in engineering education*. Ph.D. thesis, University of Colorado at Boulder.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Yun Hu, Mingxue Liao, Pin Lv, and Changwen Zheng. An empirical study of multi-domain and multi-task learning in chinese named entity recognition. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Deep Learning*, page 743–754, Berlin, Heidelberg. Springer-Verlag.

Brent K. Jesiek, Lynita K. Newswander, and Maura Borrego. 2009. Engineering education research: Discipline, community, or field? *Journal of Engineering Education*, 98(1):39–52.

Zenun Kastrati, Blend Arifaj, Arianit Lubishtani, Fitim Gashi, and Engjëll Nishliu. 2020. Aspect-based opinion mining of students' reviews on online courses. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, ICCAI '20, page 510–514, New York, NY, USA. Association for Computing Machinery.

Zenun Kastrati, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences*, 11(9):3986.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing grammatical correctness in language learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–146, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mirka Koro-Ljungberg and Elliot P Douglas. 2008. State of qualitative research in engineering education: Meta-analysis of jee articles, 2005–2006. *Journal of engineering education*, 97(2):163–175.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Wencan Luo and Diane Litman. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Lisbon, Portugal. Association for Computational Linguistics.

Wencan Luo, Fei Liu, Zitao Liu, and Diane Litman. 2016. Automatic summarization of student course feedback. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 80–85, San Diego, California. Association for Computational Linguistics.

Tamara J Moore, Aran W Glancy, Kristina M Tank, Jennifer A Kersten, Karl A Smith, and Micah S Stohlmann. 2014. A framework for quality k-12 engineering education: Research and development. *Journal of pre-college engineering education research (J-PEER)*, 4(1):2.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, Online. Association for Computational Linguistics.

Vu Duc Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2018. Variants of long short-term memory for sentiment analysis on vietnamese students' feedback corpus. In *2018 10th international conference on knowledge and systems engineering (KSE)*, pages 306–311. IEEE.

Nikola Nikolić, Olivera Grljević, and Aleksandar Kovačević. 2020. Aspect-based sentiment analysis of reviews in the domain of higher education. *The Electronic Library*.

Barbara M Olds, Barbara M Moskal, and Ronald L Miller. 2005. Assessment in engineering education: Evolution, approaches and future collaborations. *Journal of Engineering Education*, 94(1):13–25.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

John W. Prados. 1998. Engineering education in the united states: Past, present, and future.

Kevin Pugh. 2002. Teaching for transformative experiences in science: An investigation of the effectiveness of two instructional elements. *Teachers College Record*, 104:1101–1137.

Arti Ramesh, Shachi H. Kumar, James Foulds, and Lise Getoor. 2015. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 74–83, Beijing, China. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Sheri D. Sheppard, Shannon K. Gilmartin, Helen L. Chen, Krista Donaldson, Gary Lichtenstein, Ozgur Eris, Micah Lande, and George Toye. 2010. Exploring the engineering student experience: Findings from the academic pathways of people learning engineering survey (apples).

Reed Stevens, Kevin O'connor, Lari Garrison, Andrew Jocuns, and Daniel M Amos. 2008. Becoming an engineer: Toward a three dimensional view of engineering learning. *Journal of Engineering Education*, 97(3):355–368.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Charles Welch and Rada Mihalcea. 2016. Targeted sentiment to understand student comments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2471–2481, Osaka, Japan. The COLING 2016 Organizing Committee.

Miaomiao Wen, Diyi Yang, and Carolyn Rosé. 2014a. Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 525–534.

Miaomiao Wen, Diyi Yang, and Carolyn Rose. 2014b. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational data mining 2014*. Citeseer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

A Appendix

A.1 Construct Descriptions

Here, we provide a description of each construct, including when a particular affect label is annotated for the construct.

- **Expansion of Perception:** A response was tagged as Expansion of Perception if the student expressed seeing aesthetics in their daily

life. Students who expressed that aesthetics were generally unimportant were tagged as negative expansion of perception.

- **Motivated Use:** A response was coded as relating to motivated use if a student expressed a desire (or lack thereof) to use aesthetics and design in their work or daily lives. Additionally, if a student expressed that they felt that their learning could be applied, their response was tagged for motivated use.

- **Affective Value:** In order for a response to be tagged with a shift in affective value the student needed to provide an emotional response about a topic relating to those discussed in AesDes, this meant that student responses proving a positive feeling towards aesthetics or design would be flagged as experiencing a positive shift in affect.

- **Disciplinary Knowledge:** A response was tagged for Disciplinary Knowledge if the student discussed their perception of their learning. Very few students discussed Disciplinary Knowledge in their open responses, and no neutral Disciplinary Knowledge code was found.

- **Identification:** A response was tagged for identification if the student discussed either seeing themselves as an engineer, such as saying "I am an engineer" or if they mentioned someone else calling them anengineer. No students provided responses that were indicative of negative Identification.

- **Navigation:** A response was tagged for navigation if it discussed how the student felt that they were doing things that engineers do, such as accepting a position as a full-time engineer after graduation. Responses were marked as having negative navigation only if not feeling like an engineer was expressly mentioned.