Pushing Ethics Assessment Forward in Engineering: NLP-Assisted Qualitative Coding of Student Responses

Mr. Umair Shakir, Virginia Polytechnic Institute and State University
Dr. Justin L. Hess, Purdue University at West Lafayette (COE)

Dr. Justin L Hess is an assistant professor in the School of Engineering Education at Purdue University. Dr. Hess's research focuses on empathic and ethical formation in engineering education. He received his PhD from Purdue University's School of Engineering Education, as well as a Master of Science and Bachelor of Science from Purdue University's School of Civil Engineering. He is the editorial board chair for the Online Ethics Center, deputy director for research for the National Institute of Engineering Ethics, and past-division chair for the ASEE Liberal Education/Engineering and Society division.

Matthew James P.E., Virginia Polytechnic Institute and State University

Matthew James is an Associate Professor of Practice in Engineering Education at Virginia Tech, and is a registered Professional Engineer in the State of Virginia. He holds bachelors and masters degrees from Virginia Tech in Civil Engineering.

Dr. Andrew Katz, Virginia Polytechnic Institute and State University

Andrew Katz is an assistant professor in the Department of Engineering Education at Virginia Tech. He leads the Improving Decisions in Engineering Education Agents and Systems (IDEEAS) Lab, a group that uses multi-modal data to characterize, understand, a

Pushing Ethics Assessment Forward in Engineering: NLP-Assisted Qualitative Coding of Student Responses

Abstract

Recent headlines have featured large language models (LLMs), like ChatGPT, for their potential impacts throughout society. These headlines often focus on educational impacts and policies. We posit that LLMs have the potential to improve instructional approaches in engineering education. Thus, we argue that as an engineering education community, we should aim to leverage LLMs to help resolve challenges in engineering education. This study takes up one aspect of instructional design: valid assessment of students' learning outcomes in engineering ethics. In this study, we present a method for engineering educators to implement NLP in open-ended ethics assessments (here, written responses to an ethics case scenario). Grading such open-ended responses has challenges: it requires a non-trivial time commitment and attention to consistency. To mitigate these challenges, we developed an NLP approach based on open-source, transformer-based LLMs. We applied and evaluated our NLP approach for coding students' responses to an open-ended ethics case scenario in a first-year engineering course. The results showed that our NLP approach labeled 380 out of 472 sentences accurately. Conversely, only 8% (37 out of 472 responses) were inaccurately labeled. Overall, our NLP approach provides a step toward analyzing written responses to scenario-based assessments in a scalable manner. However, it is not perfect. One current downside of our NLP approach is that it requires a large upfront time investment in setting up the system. Our future work aims to lower that barrier to entry, thereby making it more accessible to a larger group of potential users.

Keywords: *engineering ethics; assessment; natural language processing*

Introduction

Engineers working in public and private sectors are making decisions that have ramifications for the present and future generations. These ramifications make it imperative for engineering students to engage with the ethical issues embedded in their work in undergraduate degree programs. Instructors often use open-ended case scenarios to prepare engineering students for ethical decision making in their work [1]. Open-ended ethics case studies or scenarios can engage students in ethical reasoning and judgment, especially when they are delivered in flexible ways and provide opportunities for students to express their views in their own words [2]. However, assessments of ethics case scenarios have their own downsides. Perhaps most notably, such grading can be time-intensive and in large course sections with multiple instructors (such as the course we study here), grading may lack consistency for both inter-grader and intra-grader assessment. We propose to help address these challenges by implementing and testing natural language processing (NLP) to assess students' written responses to an ethics scenario.

Many existing NLP tools used for the assessment of students' responses to open-ended case scenarios are established on dictionary-based methods. Their working principle is the syntactic similarity of words or counting of words in a corpus of text. This working principle renders those dictionary-based NLP tools inflexible to respond to syntactic variations in words for describing the same idea [3]–[5]. For example, consider questions that two different students might ask whilst considering stakeholder perspectives on energy issues: (i) "How do locals think about the heating problem?", or (ii) "What are residents' perspectives on the energy issue?" While these two sentences express similar ideas, they use different words. While we argue that these sentences have similar meanings (e.g., students attending to stakeholders' views on the energy problem), dictionary-based NLP tools may not be able to cluster (or identify) such sentences. Fortunately, engineering education researchers now have methods that resolve this inflexibility of the dictionary-based NLP tools by developing NLP tools based on recent, state-of-the-art, transformer-based language models (e.g., Facebook's RoBERTa [6], Google's BERT [7], and Microsoft MPNET [8]. The working principle behind such language models is exemplified by the quote, "You shall know a word by the company it keeps" [9, p. 175]. In this study, we propose an NLP approach based on transformer-based language models to help quicken assessment of students' responses to open-ended ethics case scenarios.

Study Overview

Assessing ethics learning is a challenge in engineering education. We posit that NLP can help instructors evaluate ethics learning in a time-efficient manner. Such an approach will be especially helpful when instructors have large samples of students. In doing so, at least two processes need to happen from the instructor's side. First, one must identify themes in students' responses. Second, one must then apply the relevant rubrics. This study focuses on the first process. In doing so, we will answer the following research question in this study: "What is the accuracy of the codes generated from an NLP approach that uses a transformer-based language model and a k-Nearest Neighbors matching method to qualitatively analyze students' responses to an open-ended question prompt of an ethics case scenario?"

Background and Motivation

In this section, first, we share existing ethics assessment instruments used in engineering education. Next, we summarize methods of case-based instruction in engineering ethics education literature. Finally, we discuss use of NLP in education assessment generally.

Student Outcomes and Assessment Methods in Engineering Ethics Education

For accreditation of an undergraduate engineering program, ABET has included ethics in its criteria (3-4) "an ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts" [10]. To receive ABET accreditation, engineering programs must determine and assess ethical learning outcomes[11], [12]. The constructs of ethical sensitivity, ethical knowledge, and ethical judgment are common learning goals in engineering ethics education [13]–[15].

To assess whether students developed ethics-related abilities in their engineering ethics courses, faculty members often use measurement instruments. One measure, entitled the Test of Ethical Sensitivity in Science and Engineering (TESSE), measures ethical sensitivity [16]. Another, the Survey of Engineering Ethical Development (SEED), measures ethical knowledge (e.g., knowledge of the NSPE code of ethics) [17]. Other examples include the EERI, the ESIT, DIT-2, and the Survey of Ethical Reasoning (SER) which measure ethical judgment [18]–[21]. While these quantitative measurement instruments can be useful, such measures can be challenging to implement [13]. Specifically, the measurements are (a) inflexible in that they cannot be adjusted to account for one's learning context, (b) purely quantitative and thus fail to elicit students' views in their own words, and (c) prime students to focus on certain ideas, thus activating extant schema [22] while foreclosing other possible responses. For example, if an instructor aimed to assess sensitivity to stakeholder identification in a course-based assignment, these measurements would not fulfill this need., Rather, what would be needed to assess such a specific learning objective would (or could) first involve prompting students to write or list potential stakeholders and then review responses to see whether students identified certain stakeholders without direct priming. In this sense, ethics case studies generally provide prompt students to express their views in their own words and thus can serve as a rich source of assessment data. As case studies are oft-used in engineering ethics education [23] and (we suspect) thus provide already-existing assessment data in many contexts, we describe below a case study-based instruction method in engineering ethics education.

Case-Based Instruction in Engineering Ethics Education

Case studies or case scenarios (we use these phrases interchangeably) present students with ethical dilemmas embedded in real-world contexts. These dilemmas generally do not have right or wrong answers, but rather better or worse decision outcomes for various stakeholders [2]. These shades of gray encourage students to think deeply about their values, experiences, and professional practice [14], [24]. Students reflect on scenarios with information provided in terms of news media reports, academic publications, regulatory documents, or other materials. Instructors in engineering classes often teach the code of ethics of professional organizations

such as the American Society of Civil Engineers and the National Society of Professional Engineers [25].

To assess students' ethical decision-making using ethics case scenarios, instructors can design and embed question prompts related to recognized ethical issues, affected stakeholders, and the various impacts on those stakeholders. Through their written responses, students make decisions that may have positive or negative implications for the people or groups involved in the case study. Students are expected to support the logic of their decisions by professional code of ethics and ethical decision making theories. Examples of those are deontology, utilitarianism, and consequentialism.

Natural Language Processing (NLP)

Natural languages are languages that evolve through human use over time as opposed to formal languages such as mathematics or computer languages. NLP is a field at the interface of human and computer languages and it focuses on how to program a computer to process human language data effectively and efficiently [26], [27]. NLP has been used to facilitate open-ended assessments of student learning, often based on the rationale that NLP can save time and effort for graders or prevent biases across multiple graders [28]–[30]. The working principle behind our use of NLP in open-ended assessment for this paper is that one can grade students' answers if (a) one has a bank of reference graded answers and (b) a way to inspect similarity of those new answers to the reference graded answers.

To examine that textual similarity, some of the earliest NLP tools were based on dictionary- or word frequency-based approaches such as bag of words (BOW). In a BOW model, the central feature of textual similarity is lexical similarity (or same wording) only. For example, when using BOW, 'Adam is heavier than John' is identical to 'John is heavier than Adam'. Those earlier NLP tools failed to account for feature of semantic or syntactic similarity (i.e., relations between words in sentences). To address this challenge of capturing richer semantic meaning in text, in the early 2010s, a breakthrough in the field of NLP was achieved by the development of word embeddings methods such as Word2Vec [4], [5].

Word embeddings are intended to be high-dimensional abstract representations of words or phrases in a vector space. A less mathematical way of stating this is that we want to try and represent each sentence with a long array of numbers to retain semantic features of words in a sentence. What each of those numbers in that array means individually is not particularly important. The key is in how each numerical representations of the sentences relates to each other. In theory, one would want similar words to have similar representations. For example, vector representation of 'fantasy' and 'imagination' would have similar angles in a high-dimensional space since their semantic meaning is similar. In word embedding techniques, the similarity between sentences or phrases is measured by calculating the distance between their vector representations. The common examples of those distances are cosine distance, Euclidian distance, or Manhattan distance. The word embedding models contributed substantially to developing pre-trained large language models (LLMs) using corpora such as Wikipedia or university library repositories.

LLMs are state-of-the-art NLP tools and can be finetuned on further downstream NLP tasks. LLMs are developed on neural networks machine learning architectures—a.k.a., transformers architectures—which enable the models to effectively learn long-dependencies in sequences of phrases or sentences and extract semantic context [27], [31]. The ability to effectively extract semantic context is the most relevant feature of LLMs in comparing unlabeled students' responses to example responses. Promising LLMs include BERT [7], Generative Pretraining Transformer (GPT-3) [32], and Masked and MPNet [8]. Our NLP approach uses the open-source MPNet transformer language model available from the Hugging Face repository.

Study Design

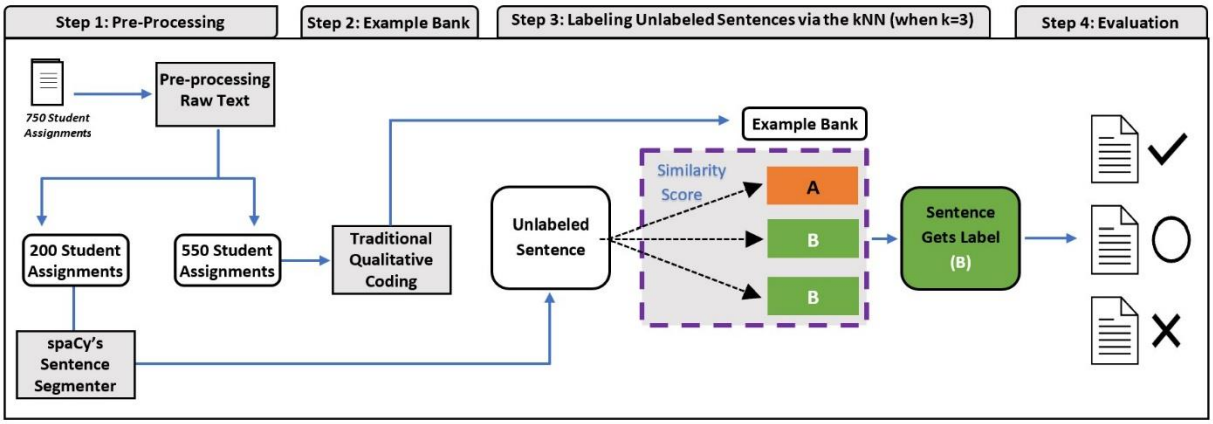
The research process of this study comprised four steps as shown in Figure 1. An overview of these four steps is provided here, followed by more in-depth discussion in subsequent sections.

Figure 1. An Overview of Study Design

First, we collected data (750 students' assignments) from instructors and pre-processed the raw text data before passing it to the NLP workflow. Second, we took a subset of 550 students' assignments to do traditional qualitative coding for developing an example bank. Third, we assigned labels to the unlabeled remaining subset of 200 students' assignments with the NLP approach. Lastly, we read those (newly) labeled students' responses to evaluate whether assigned codes to those responses through the NLP approaches were accurate or not. Here, accuracy means that the assigned code represented the idea expressed in student responses. We technically implemented those four processes in Google Colab notebooks that were written using a combination of the R and Python programming languages. All code is presented in the GitHub repository we have set up for this project at: https://github.com/andrewskatz.

Data Collection

The first-year engineering program (FYE) at Virginia Tech teaches students an ethics module that comprises a case-based instructional design of two hours in a semester. While there are several ethics cases that instructors may use in the FYE, the most commonly used case is the Big Belly Trash Can. For assessment purposes of the ethics module, the students are required to



* White represents data (e.g., students' responses)
** Gray represents process

submit their written responses to question prompts of the Big Belly ethics case scenario. These question prompts are related to (a) recognition of an ethical issue, (b) identification of a stakeholder, (c) possible decision choices according to various ethical decision-making theories, and (d) consequences of those decisions on various stakeholders. We include the case scenario and question prompts in Appendix A. We collected written responses to the ethics case scenario from 750 students who consented to participate in research associated with the class.

The original data was generated as students' assignments but not with the explicit purpose of being used for research purposes in NLP. Therefore, some of the question prompts of the case scenario were phrased in a suboptimal manner for the NLP approaches. This is because students may describe multiple ideas in a single short sentence, which can sometimes lead to noise in the NLP automated analysis. These approaches work best when the respondent focuses on one idea at a time. This challenge of eliciting multiple pieces of information at once due to the question phrasing and response format is a limitation of our approach. For this study, we used students' responses to the question prompt: "identify an ethical dilemma or issue from the case study." We chose this question prompt because students' responses to that tend to be more structured and focus on one idea at a time. Therefore, this question prompt presented an opportunity to demonstrate how NLP approaches can work for qualitative coding of students' responses.

Pre-Processing of Raw Text Data

We collected 750 students' assignments as pdf files from instructors. After converting the pdf files to text files, we removed Arabic numerals such as "[1], [2], etc.", from the excerpts (note: students were instructed to follow the IEEE citation style, but citations themselves were not the goal of our analysis). From the total of 750 students' assignments, we randomly separated the data into two subsets of 550 and 200 students' assignments. We performed this split in order to have some data (n = 550)for developing the codebook and model and a separate set (n = 200) student assignments for the actual evaluation of the model performance.

Data Analysis

Our analysis involved three steps: (i) developing example bank, (ii) labeling students' responses, and (iii) evaluating assigned labels.

Developing Example Bank via Traditional Qualitative Coding

We developed an example bank by qualitatively coding a subset of 550 students' written responses to select question prompts. Here, the purpose was to develop a codebook that covers all possible aspects of responses to the question prompt, and the labels were assigned to those responses. We first uploaded 550 (pre-processed) students' written responses in the traditional qualitative coding software Dedoose. We next read students' responses to define, refine, and assign codes [33]. When we observed the saturation point was reached (i.e., no new codes were emerging), we downloaded all codes and their excerpts from Dedoose in .csv format. This file comprised our initial example bank.

The most commonly identified ethical issue was whether or not to install the Big Belly trash cans in Sans Francisco. Students saw this as a tradeoff between keeping the city clean and removing a source of income for the homeless population in the city. The second most common ethical dilemma related to data privacy concerns. This involved a tradeoff between sending digital information to waste management workers and collecting user information. Smaller groups of students identified other key ethical concerns, including (1) loss of food source for wildlife as a result of reducing waste as a food source and (2) loss of jobs for current waste management workers as a result of installing trash can those have more capacity. Table 1 lists a few of the different ways that students discussed these ethical issues and potential responses.

Table 1. A Snapshot of Example Bank

(Parsed) Student Responses	Assigned Label
An ethical dilemma that arises with this case study is that when these new trash compactor garbage cans are installed, it removes that source of income for the homeless community that took advantage of the loose garbage. An ethical dilemma that emerged from the case study involving the Big Belly Solar trash cans is that the homeless population in surrounding areas are no longer able to collect cans and bottles from the open trash cans Big Belly Solar trash cans or not, their solution will most likely cut off a reliable source of income for the homeless Big Belly ultimately changes the lives of the homeless and takes away their one source of income	Access to Income
An ethical dilemma the waste bins face is maintaining public safety/ privacy. With the addition of smart trash cans, it is able to collect data with the use of sensors and cameras. Big Belly CEO Jack Kutner proposed adding data collection devices onto these waste disposal receptacles, and I think that an action like this would involve ethical concerns relating to privacy issues. When reading the case study, an issue that stood out to me was that the CEO wanted to make the trash cans capable of monitoring a good amount of its surroundings. This includes “temperature, bin usage, foot traffic on the street, humidity and other important data”(Atkinson). Delving in further into the company and the CEO’s interests, it is concerning what the future holds for these trash can	Privacy Concerns

Labeling the Unlabeled Student Responses via k-Nearest Neighbors (kNN)

To thematically analyze the remaining 200 students' assignments, they were matched to response excerpts in the example bank through the k Nearest Neighbors (kNN) method [34]. Here, a noteworthy process is that we split 200 students' responses at sentence level via spaCy's sentence segmenter [35] before passing those to the kNN classifier for thematic analysis. The spaCy's sentence segmenter yielded 912 response sentences. We did this to capture more of the nuance in what students wrote. Students' responses to the question prompt were typically a single block of text, which may or may not have consisted of multiple sentences (and multiple themes). A block of text (e.g., a paragraph or a sentence) might express multiple topics at a time, but the vast majority of single phrases (or sentences) express only one topic. Next, we describe the technical implementation of the kNN method.

The technical implementation of kNN method included the following three steps: (a) sentence embedding, (b) calculating cosine similarity, and (c) assigning labels by identifying a majority vote (when k =3).

Sentence Embedding. We embedded the raw text—sentences from the unlabeled dataset and the example bank—into a 768-dimensional vector space using the pre-trained MPNet embedding model [8]. The pre-training means the model have been already trained on large text corpus to generate embeddings. After embedding both data sets, we determined similarity scores between unlabeled sentences and labeled sentences.

Similarity Score. We used the cosine similarity score between embedding vectors of labeled and unlabeled sentences [36]. Theoretically, the similarity score will range from 0 to 1. The maximum value similarity score is 1, which represents the exact match between unlabeled and labeled sentences. As the similarity score between two sentences decreases from 1 to 0, we infer that those two sentences do not match each other—in other words, they are less likely to be about the same topic. Each unlabeled sentence will have a similarity score from its comparison with each sentence from the example bank as shown in Figure 1. The question is which label of an example bank sentence should be assigned to an unlabeled sentence. To achieve this purpose, we used the kNN method of majority vote (when k =3).

Assigning labels (when k = 3). First, we selected three example bank sentences with the highest similarity scores with an unlabeled sentence. Among labels of those three example sentences, any label with a majority vote (2 or more) was assigned to the unlabeled sentence. For example, in Figure 1, the unlabeled sentence is assigned label B because it has the majority of 2. If any label does not have the majority vote (2 or more), the unlabeled sentence will remain unlabeled. For example, if a sentence of the example bank in Table 1 has the label C rather than B, the unlabeled sentence will not be assigned any label among the three (A, B, C) because none of the labels has a majority. In this study, we input 912 unlabeled response sentences to the kNN matching method. Among those 912 response sentences, 440 (45%) remained unlabeled and 472 (55%) were assigned labels. Taking these 472 response sentences with their assigned labels, we performed the following evaluation procedure to answer our RQ.

Evaluation Procedure

We read each sentence or phrase to evaluate whether the assigned code represented the idea described in the sentence. If yes, then we assigned it a rating of an accurate label as 1. If the assigned code did not match our qualitative coding, then we assigned it a rating of an inaccurate label as -1. In between those extreme ratings, we have a third category of neutral as 0. We used this category in instances of ambiguity or partial credit. For example, a sentence could be about more than one idea or the sentence itself might be ambiguous. Lastly, we used numerical evaluation ratings to calculate the total number (and percentages) of sentences that are labeled (a) accurately, (b) inaccurately, and (c) neutral by the NLP approach. This quantitative evaluation procedure allowed us to answer this study's RQ.

Limitations

There were several limitations of this study. The first limitation is specific to this paper rather than the study itself: the examples in the results table shown below were selectively chosen to illustrate the output of our NLP approach. The actual output was too unwieldy to provide here because there were too many sentences and their assigned labels. The second limitation is an algorithmic one: there is no guarantee that all response sentences will get a label in the kNN method. This is because our NLP workflow is set up in such a way that sentences were only labeled (and appear in final output) when the label had a majority vote of two (i.e., when $k = 3$). A third limitation of this study is related to the step of splitting sentences. This step is a tradeoff between accuracy and utility. Regarding accuracy, splitting sentences can lead to losing the context of what a student is saying when students refer to a previous sentence, like a pronoun with an ambiguous referent. For example, consider: "The students read the case studies. They liked them." When we split these two sentences, one cannot understand whom "they" and "them" refer to. On the other hand, regarding utility, a block of text (e.g., a paragraph) without splitting sentences might express multiple ideas at a time. This may lead to suboptimal performance for the NLP approach. We have also elaborated on the tradeoff between accuracy and utility in the data collection section. A fourth limitation is that we used convenience sampling (students' assignments) and therefore could not modify the data collection procedure. The assignment was given as a regular part of the course without the intention of being used in research related to NLP. Because of their relevance, availability, and their large volume (more responses tend to result in better outcomes, based on our experience), we leveraged those student assignments here for demonstration purposes of our NLP approach. Despite the above limitations, we argue this study has novel utility for ethics education community members, especially those who strive to scale open-ended ethics assessments in their large classrooms.

Results and Discussion

Table 2 presents the quantitative results of the evaluation procedure we used to address our research question, "What is the accuracy of the codes generated from an NLP approach that uses a transformer-based language model and a k-Nearest Neighbors matching method to qualitatively analyze students' responses to an open-ended question prompt of an ethics case scenario?" Not counting 45% of the dataset which was unlabeled ($n = 440$), the accuracy of our NLP approach was 81%, as it accurately labeled 380 out of 472 response sentences. The study's

NLP approach inaccurately labeled 8% of the response sentences, or 37 out of 472. The remaining 55 (11%) responses sentences received a partial credit rating during our evaluation procedure. These 55 responses sentences were ambiguous or represented multiple topics in a short sentence. Two examples of the ambiguous sentences for the purpose of our analysis included: “The main ethical dilemma/issue is, is it fair?”, “I have always grown up caring about the environment.”

Table 2. Evaluation Rating of Sentences Labeled by the NLP Approach

Description of Evaluation Rating	Counts	Percentage (%)
Accurately Labeled (1)	380	81
Partial Credit or Ambiguous Sentence (0)	55	11
Inaccurately Labeled (-1)	37	7.8
	472	**100**

In Table 3 we provide three (example) response sentences that were accurately assigned labels with our NLP approach: access to income and environment versus homeless people. These examples demonstrated how the NLP approach successfully labeled syntactically different but semantically similar student responses. For instance, in the case of the “access to income” label, a student wrote about the ethical dilemma as “rejection of a source of income for the homeless”. Another student wrote about the same ethical dilemma in different words: “income to be taken away from the homeless”. The NLP approach based on transformers assigned the same label to those two example students’ responses having different words but expressing the same idea. This flexibility of the NLP approach is promising for engineering educators. With our NLP approach, instructors and scholars can qualitatively analyze students’ responses to scenario-based assessments without a human (team) to read all of those responses.

Table 3. Examples for Evaluation Ratings for Accurately Labeled Sentences

(Parsed) Student Responses (note: bold emphases were added by our team)	Assigned Label
I believe that the biggest ethical dilemma that is presented into this case study is **the rejection of a source of income** for the homeless that surround the urban areas around the bay.	Access to Income

By implementing the big belly solar trash compactor system, there is less likely going to be as many bottles littered cause **this income to be taken away from the homeless** One ethical dilemma in this case study is the fact that these **barrels prevent the homeless from collecting bottles and cans to turn in for money, which tends to be their only source of income.**	
One of the ethical dilemmas from this case study is the fact that these new Big Belly trash cans, while **beneficial for the environment and looks of the campus, have the potential to harm the homeless population** due to the fact that it makes it impossible to dig through the trash for cans. Here, we must determine **whether the needs of these homeless people outweigh the needs of others for a cleaner environment and a cleaner city.** Additionally, due to the fact that homeless people use the trash to gain **a source of income, removing that source by tidying up provides the other side of the dilemma.**	Environment versus homeless

Instead of the unsupervised machine learning (kNN) we presented here, many NLP researchers also used supervised machine learning approach. In recently published studies, [37], [38] trained linear regression models for matching the written answers of their study's participants to assign code. Those authors reported that their models yield qualitative coding of students' answers with the same level of inter-rater reliability between computer and human as that between two humans performing traditional qualitative coding of students' answers. However, we argue that our NLP has more utility for the engineering ethics education community than supervised learning models but we emphasize that one challenge is the training of the models. The main idea behind that training is to provide an algorithm with a labeled dataset to learn key features of the input dataset that would produce the appropriate output (i.e., labels). This training process requires more data and a higher level of technical familiarity with coding and/or collaboration with individuals more familiar with the training process. Yet, this up-front investment in training and time can yield more parity across graders and facilitate grading for others, including those untrained in the dataset. Thus, we believe that our NLP approach offers greater utility for the ethics education community compared to supervised learning models.

Conclusions and Future Work

The engineering education community should aim to leverage large Language Models (LLMs) to help resolve challenges in engineering education, such as valid assessment of students' learning outcomes. In this study, we presented an NLP approach based on open-source, transformer-based LLMs for engineering educators to implement while assessing open-ended responses to case

studies. In that spirit, we applied and evaluated our NLP approach for qualitative coding of students' responses to an open-ended ethics case scenario in a first-year engineering course. The results showed that our NLP approach labeled 380 out of 472 sentences accurately. We acknowledge the limitations of our NLP approach, such as that it requires large upfront time investment in setting up the system. Our future work aims to lower such barriers to entry, thereby making it more accessible to a larger group of users. Another dimension of our future work aims to extend this work for grading written responses. The philosophy behind this is that once the space of possible answers has been saturated, representatives of each kind of response can be included in an answer bank. This answer bank can then be used for labeling new responses from students by comparing each new response to the pre-labeled responses in the answer bank. If the labels in the answer bank also include a score associated with that response, then one could append that score to the matched new response. Such use of NLP can lead to greater inter-grader and inter-grader consistency in large course sections with multiple instructors and graders, which, we argue, has the potential to promote fairness and equality in engineering ethics assessment processes.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 2107008. We also would like to thank the students whose responses we analyzed in this study and the instructors who agreed to support our investigation.

Appendix A

Big Belly Solar Case Study

Background

The problem of waste management in urban settings is a problem that cities have been working to tackle for a long time. Recently a number of new technologies, developed in part by engineers, have emerged to help combat common trash problems. The Big Belly Solar trash compactor system is one of the technologies that have been widely implemented, including on our own campus here at Virginia Tech. As with many new technologies, there is some controversy about whether these types of trash cans should be adopted widely, with arguments on either side. The cases that you will read about look at two perspectives of the Big Belly Solar roll out in the San Francisco Bay area--one in the City of San Francisco, and another across the bay at the University of California, Berkeley. The third article is more recent and elaborates on San Francisco's most recent efforts to prototype their own trash bin.

Before you read these three articles, it is important to understand some context that differs from what you may be familiar with. In contrast to the relatively rural setting that most of you are familiar with here at Virginia Tech, UC Berkeley is located in an urban environment. This means that some of the challenges found in the surrounding community, like homelessness, are more visible on their campus. It is also important to note that in both Berkeley and San Francisco, unlike Virginia, California offers a container deposit incentive such that someone can turn in used containers for 5 or 10 cents each. It is not uncommon for homeless people in the state to work to collect discarded bottles and make money from returning them as a source of income.

Case study articles

The following three articles present unique views of the implementation of the Big Belly solar trash cans, and cover some of the successes and challenges encountered. Read through the articles, and use this information to help you complete your ethics warm-up in class 10B, as well as the individual ethics report.

As you read through these articles, think about some of the stakeholders that are either directly or indirectly involved. If you have trouble identifying a stakeholder, you can pick something from the list at the end of this document. Also, think about the role that engineers might have taken on in each of the cases.

Berkeley Big Belly Solar Case

Big Belly Solar: Sproul's New Waste Bins

San Francisco Big Belly Solar Case

Talkin' trash: Little appetite in San Francisco for Big Belly garbage bins

San Francisco Trash Bin Prototyping Efforts

Garbage odyssey: San Francisco's bizarre, costly quest for the perfect trash can

Prompts

Analyze the given ethical case study and materials provided in the assignment N6 description, and answer the following questions. Each answer should be in full sentences, ~1 paragraph.

1. Identify an ethical dilemma or issue from the case study. (Note that you are welcome to infer more details of settings to the case study or make reasonable assumptions about what is going on in order to answer this question and the next one.)

2. Explain why the above answer is an ethical dilemma/issue.

3. Describe at least two ethical theories that could be applied to help make sense of the ethical dilemma/issue you described. Explain *how* these two ethical theories are relevant and would be applied to the ethical dilemma/issue you described. (You can find details of the four ethical theories with links to videos in the N6 assignment description document)

4. Describe a stakeholder related to this ethical dilemma/issue. Explain why you chose them and how they are related to the ethical dilemma/issue. (Note that accounts of your stakeholder shouldn't be limited by our text of the case study. You are welcome to create more details for your stakeholders or make reasonable assumptions about them in order to answer this question and the next one.)

5. Explain the differences in the impact on your chosen stakeholder between the two ethical theories that you described above. For example, what would the consequence be for your stakeholder if an engineer or decision maker applied one framework versus the other?

6. List one fundamental canon of the NSPE code of ethics that applies to this situation, and explain *how* it is related.

References

[1] J. L. Hess and G. Fore, "A Systematic Literature Review of US Engineering Ethics Interventions," *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 551–583, Apr. 2018, doi: 10.1007/s11948-017-9910-6.

[2] D. A. Martin, E. Conlon, and B. Bowe, "Using case studies in engineering ethics education: the case for immersive scenarios through stakeholder engagement and real life data," *Australas. J. Eng. Educ.*, vol. 26, no. 1, pp. 47–63, Jan. 2021, doi: 10.1080/22054952.2021.1914297.

[3] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing." arXiv, Aug. 28, 2021. doi: 10.48550/arXiv.2108.05542.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." arXiv, Sep. 06, 2013. doi: 10.48550/arXiv.1301.3781.

[5] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in Pre-Training Distributed Word Representations." arXiv, Dec. 26, 2017. doi: 10.48550/arXiv.1712.09405.

[6] Y. Liu *et al.*, "Roberta: A robustly optimized BERT pretraining approach." arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[8] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 16857–16867. Accessed: Feb. 10, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html

[9] J. Firth, "Descriptive linguistics and the study of English," *World Englishes Crit. Concepts Linguist. Ed K Bolton B Kachru*, vol. 3, pp. 203–217, 1968.

[10] ABET, "Criteria for Accrediting Engineering Programs, 2022 – 2023," 2023. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2022-2023/ (accessed Feb. 10, 2023).

[11] B. E. Barry and M. W. Ohland, "ABET criterion 3.f: How much curriculum content is enough?," *Sci. Eng. Ethics*, vol. 18, no. 2, pp. 369–392, Jun. 2012, doi: 10.1007/s11948-011-9255-5.

[12] N. E. Canney, M. Polmear, A. R. Bielefeldt, D. Knight, C. Swan, and E. Simon, "Challenges and Opportunities: Faculty Views on the State of Macroethical Education in Engineering," presented at the 2017 ASEE Annual Conference & Exposition, Columbus, Ohio, Jun. 2017. Accessed: Mar. 16, 2020. [Online]. Available: https://peer.asee.org/challenges-and-opportunities-faculty-views-on-the-state-of-macroethical-education-in-engineering

[13] M. Davis and A. Feinerman, "Assessing Graduate Student Progress in Engineering Ethics," *Sci. Eng. Ethics*, vol. 18, no. 2, pp. 351–367, Jun. 2012, doi: 10.1007/s11948-010-9250-2.

[14] B. Newberry, "The dilemma of ethics in engineering education," *Sci. Eng. Ethics*, vol. 10, no. 2, pp. 343–351, 2004, doi: 10.1007/s11948-004-0030-8.

[15] N. H. Steneck, “Designing teaching and assessment tools for an integrated engineering ethics curriculum,” in *FIE’99 Frontiers in Education. 29th Annual Frontiers in Education Conference. Designing the Future of Science and Engineering Education. Conference Proceedings (IEEE Cat. No.99CH37011*, Nov. 1999, p. 12D6/11-12D6/17 vol.2. doi: 10.1109/FIE.1999.841685.
[16] J. Borenstein, M. Drake, R. Kirkman, and J. Swann, “The Test of Ethical Sensitivity in Science and Engineering (TESSE): A Discipline Specific Assessment Tool for Awareness of Ethical Issues,” presented at the 2008 Annual Conference & Exposition, Jun. 2008, p. 13.1270.1-13.1270.10. Accessed: May 19, 2022. [Online]. Available: https://peer.asee.org/the-test-of-ethical-sensitivity-in-science-and-engineering-tesse-a-discipline-specific-assessment-tool-for-awareness-of-ethical-issues
[17] C. J. Finelli *et al.*, “An Assessment of Engineering Students’ Curricular and Co-Curricular Experiences and Their Ethical Development,” *J. Eng. Educ.*, vol. 101, no. 3, pp. 469–494, 2012, doi: 10.1002/j.2168-9830.2012.tb00058.x.
[18] J. Borenstein, M. J. Drake, R. Kirkman, and J. L. Swann, “The Engineering and Science Issues Test (ESIT): A Discipline-Specific Approach to Assessing Moral Judgment,” *Sci. Eng. Ethics*, vol. 16, no. 2, pp. 387–407, Jun. 2010, doi: 10.1007/s11948-009-9148-z.
[19] H. Clarkeburn, “A Test for Ethical Sensitivity in Science,” *J. Moral Educ.*, vol. 31, no. 4, pp. 439–453, Dec. 2002, doi: 10.1080/0305724022000029662.
[20] P. W. Odom and C. B. Zoltowski, “Statistical Analysis and Report on Scale Validation Results for the Engineering Ethical Reasoning Instrument (EERI),” presented at the 2019 ASEE Annual Conference & Exposition, Jun. 2019. Accessed: May 19, 2022. [Online]. Available: https://peer.asee.org/statistical-analysis-and-report-on-scale-validation-results-for-the-engineering-ethical-reasoning-instrument-eeri
[21] J. R. Rest, *Moral development: Advances in research and theory*. New York: Praeger, 1986.
[22] J. R. Rest, D. Narvaez, S. J. Thoma, and M. J. Bebeau, “DIT2: Devising and testing a revised instrument of moral judgment,” *J. Educ. Psychol.*, vol. 91, pp. 644–659, 1999, doi: 10.1037/0022-0663.91.4.644.
[23] J. L. Hess and G. Fore, “A Systematic Literature Review of US Engineering Ethics Interventions,” *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 551–583, Apr. 2018, doi: 10.1007/s11948-017-9910-6.
[24] J. R. Herkert, “Engineering ethics education in the USA: Content, pedagogy and curriculum,” *Eur. J. Eng. Educ.*, vol. 25, no. 4, pp. 303–313, 2000, doi: 10.1080/03043790050200340.
[25] J. Sliwa, “Ethics between the lines (of code),” in *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering*, May 2014, pp. 1–7. doi: 10.1109/ETHICS.2014.6893460.
[26] M. Edalati, “The Potential of Machine Learning and NLP for Handling Students’ Feedback (A Short Survey).” arXiv, Nov. 07, 2020. doi: 10.48550/arXiv.2011.05806.
[27] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, “Progress in Neural NLP: Modeling, Learning, and Reasoning,” *Engineering*, vol. 6, no. 3, pp. 275–290, Mar. 2020, doi: 10.1016/j.eng.2019.12.014.
[28] S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio, “Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers.” arXiv, Mar. 11, 2022. Accessed: Aug. 06, 2022. [Online]. Available: http://arxiv.org/abs/2204.03503

[29] A. I. Aldea, S. M. Haller, and M. G. Luttikhuis, “Towards Grading Automation of Open Questions Using Machine Learning,” presented at the 48th SEFI Annual Conference on Engineering Education, 2020, pp. 584–593.
[30] M. Ahmad, K. Junus, and H. B. Santoso, “Automatic content analysis of asynchronous discussion forum transcripts: A systematic literature review,” *Educ. Inf. Technol.*, vol. 27, no. 8, pp. 11355–11410, Sep. 2022, doi: 10.1007/s10639-022-11065-w.
[31] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jan. 29, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
[32] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, “What Makes Good In-Context Examples for GPT-3?” arXiv, Jan. 17, 2021. doi: 10.48550/arXiv.2101.06804.
[33] V. Clarke and V. Braun, “Thematic analysis,” *J. Posit. Psychol.*, vol. 12, no. 3, pp. 297–298, May 2017, doi: 10.1080/17439760.2016.1262613.
[34] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, p. 43:1-43:19, Jan. 2017, doi: 10.1145/2990508.
[35] M. Honnibal and I. Montani, “Spacy,” *Nat. Lang. Underst. Bloom Embed. Convolutional Neural Netw. Increm. Parsing*, 2017.
[36] M. Putnikovic and J. Jovanovic, “Embeddings for Automatic Short Answer Grading: A Scoping Review,” *IEEE Trans. Learn. Technol.*, pp. 1–13, 2023, doi: 10.1109/TLT.2023.3253071.
[37] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, “Classification of Open-ended Responses to a Research-based Assessment Using Natural Language Processing,” *Phys. Rev. Phys. Educ. Res.*, vol. 18, no. 1, p. 010141, Jun. 2022, doi: 10.1103/PhysRevPhysEducRes.18.010141.
[38] F. Ahmad *et al.*, “A Deep Learning Architecture for Psychometric Natural Language Processing,” *ACM Trans. Inf. Syst.*, vol. 38, no. 1, p. 6:1-6:29, Feb. 2020, doi: 10.1145/3365211.