Applying natural language processing to automatically assess student conceptual understanding from textual responses

Rick Somers, Samuel Cunningham-Nelson, Wageeh Boles
Queensland University of Technology

In this study, we applied natural language processing (NLP) techniques, within an educational environment, to evaluate their usefulness for automated assessment of students' conceptual understanding from their short answer responses. Assessing understanding provides insight into and feedback on students' conceptual understanding, which is often overlooked in automated grading. Students and educators benefit from automated formative assessment, especially in online education and large cohorts, by providing insights into conceptual understanding as and when required. We selected the ELECTRA-small, RoBERTa-base, XLNet-base and ALBERT-base-v2 NLP machine learning models to determine the free-text validity of students' justification and the level of confidence in their responses. These two pieces of information provide key insights into students' conceptual understanding and the nature of their understanding. We developed a free-text validity ensemble using high performance NLP models to assess the validity of students' justification with accuracies ranging from 91.46% to 98.66%. In addition, we proposed a general, non-question-specific confidence-in-response model that can categorise a response as high or low confidence with accuracies ranging from 93.07% to 99.46%. With the strong performance of these models being applicable to small data sets, there is a great opportunity for educators to implement these techniques within their own classes.

Implications for practice or policy:

- Students' conceptual understanding can be accurately and automatically extracted from their short answer responses using NLP to assess the level and nature of their understanding.
- Educators and students can receive feedback on conceptual understanding as and when required through the automated assessment of conceptual understanding, without the overhead of traditional formative assessment.
- Educators can implement accurate automated assessment of conceptual understanding models with fewer than 100 student responses for their short response questions.

Keywords: natural language processing (NLP), automated assessment of understanding, formative assessment, machine learning, conceptual understanding, mixed methods

Introduction

Assessing students' conceptual understanding and providing timely feedback are crucial aspects of teaching as these allow for teaching to be tailored to better develop conceptual understanding efficiently. With the emerging shift in favour of flexible study arrangements, traditional formative assessment techniques are less suitable and applicable in today's teaching (Gikandi et al., 2011). Automated assessment allows feedback to be delivered to educators and students as and when required in a reproducible manner. Furthermore, with an automated approach, the detrimental effects of increasing class sizes are significantly reduced with the time requirements of traditional assessment being drastically reduced.

There exists a plethora of potential applications for natural language processing (NLP) in education, such as dialogue-based tutoring systems, paraphrasing tools and text quality software (Burstein, 2009). Examples of NLP applications are educational chatbots (Kerly et al., 2007), automatic grading systems (Smith et al., 2020) and tools for tracking educational experiences (Denny et al., 2009). However, there have been few examples of NLP applied to assess a deeper level of understanding. Therefore, this study aimed to complement existing works of NLP applications in education, by investigating their potential in automatically assessing students' conceptual understanding.

Automated assessment solutions have primarily focused on the accurate marking and grading of students' work. Automated understanding assessment differs in that it provides students with the opportunity to receive accurate feedback into their conceptual understanding, allowing students to self-assess and review their knowledge when they desire, without restrictions due to educator availability and time. The ability to review conceptual understanding presents additional benefits to students; they can identify and address misconceptions and confirm their understanding on concepts rapidly, enabling them to build confidence in their understanding.

By reducing the time dedicated to traditional formative assessment, educators reap additional benefits by implementing automated assessment of students' conceptual understanding as part of their teaching. Educators can dedicate more time to teaching, developing their teaching practices and resources and addressing students' questions and uncertainties. With feedback on a cohort's conceptual understanding, educators have the opportunity to tailor their teaching to most effectively build their students' conceptual understanding while addressing misconceptions. This feedback also presents opportunities for educators to reflect and improve upon their teaching practices and resources, for both their current and future classes.

For students, formative assessment provides valuable feedback as a means of reviewing their conceptual understanding. With larger cohorts, the amount of time educators can dedicate to accurately assessing and providing useful feedback to students is restricted (McCarthy, 2017). Furthermore, with the reduction in face-to-face time in online education, there are fewer opportunities for traditional formative assessment that provides students with timely feedback on their conceptual understanding. NLP provides an opportunity to extract key insights into students' conceptual understanding allowing for an automated assessment approach.

Literature review

Assessment of conceptual understanding

Assessment of a person's conceptual understanding can be realised by evaluating evidence of their ability to transfer their knowledge and skills to new situations and scenarios (Wiggins & McTighe, 2005). This evidence can be found through appropriate assessment in a classroom environment. To assess conceptual understanding, assessments must be designed in a way that evidence of transferability can be discerned. Therefore, the design of assessment is paramount in its ability to provide this evidence.

Formative assessment is valuable in its ability to provide feedback to both educators and students to guide decisions to achieve learning outcomes (Dodge, n.d.). Through formative assessment, educators can discern evidence of students' skills, knowledge and conceptual understanding. Passing these insights onto the students provides them with an opportunity to guide their self-learning. The feedback students receive is valuable as people are often poor at accurately judging what they do and do not know (List & Alexander, 2015).

With evidence of students' skills, knowledge and conceptual understanding, teachers can better achieve learning outcomes by implementing a constructivist teaching approach (Keeley, 2008). By evaluating preconceptions on topics, learning can be targeted to build and develop these effectively. Evaluating students' current conceptions can provide insight into the accuracy of their knowledge, skills and conceptual understanding and can be used to evaluate whether misconceptions have been or are being developed.

Automated assessment approaches

Many automated formative and summative assessment techniques have been developed and adopted by educational institutes. However, due to their nature, most cannot be applied to assess conceptual understanding; those that do have often lacked in accuracy or their ability to provide insight into the nature of students' conceptual understanding.

Multiple-choice style assessments are widely used due to their objective nature and efficient marking, which can be easily automated to provide fast feedback to students and educators (Survey Anyplace, n.d.). To overcome the inherent disadvantages of this assessment type with regards to assessing conceptual

understanding, careful consideration must be given to the questions and available answers. Concept inventories have been developed to measure students' understanding of specific concepts using multiple-choice style questions (D'Avanzo, 2008; Hestenes et al., 1992; Madsen et al., 2017). By combining effective questioning and carefully designed distractors, these tests can assess students' understanding and identify misconceptions. Despite this, they are lacking in their ability to evaluate guessed selections, provide evidence for the cause or reasoning behind students' misconceptions or provide insight into the nature of students' understanding (Goncher & Boles, 2017).

Computer software has been extensively used to automatically mark short- and long-text responses. There exist several autograding approaches that can grade textual responses in a similar manner to educators, such as semantic and graph alignment features (Krithika & Narayanan, 2015) and text similarity combined with grading-specific constructs (Sultan et al., 2016); however, these approaches do not necessarily provide insight into conceptual understanding. A shortcoming in many of these is that they use text-similarity approaches to grading, comparing a student's response to a model. Therefore, they are often limited in applications where there are multiple correct responses or when different responses should be classified into the same category. Overall, the greatest drawback of autograders in terms of assessing conceptual understanding is that they have not been designed to do so; they grade textual responses, they do not assess conceptual understanding or discern evidence of understanding from students' responses.

Extraction of meaning from text through NLP

As computers cannot directly understand text, they are incapable of drawing meaning from it (Garbade, 2018). Hence, the NLP process aims to transform text into an interpretable, numerical representation. The challenge of the NLP process, specifically in the education space, is maintaining the semantic meaning of the original text in its numerical representation. The NLP process structure (see Figure 1) details the overall steps of NLP in real-world applications.

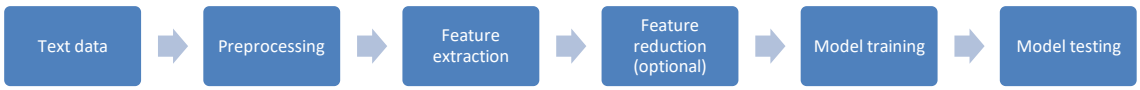

Figure 1. The NLP process structure (adapted from Cunningham-Nelson, 2019, p. 40)

The preprocessing stage aims to make the text more predictable and analysable (Ganesan, 2019). It is common practice in most NLP applications to perform lower-casing and punctuation removal as preprocessing. Additional techniques such as stemming, lemmatisation and text-enrichment may be beneficial, dependant on the application and amount of text data available.

The preprocessed text is transformed into numerical data through feature extraction (Kowsari et al., 2019). Text features can be as simple as the number of words in the text, or more complex, such as vector representations of the words (Alammar, 2019). Depending on the feature extraction techniques used, feature reduction may be beneficial. The aim on feature reduction is to reduce the size of the numerical data to make it more interpretable (Widmann & Silipo, 2015). Broadly, feature reduction techniques achieve this by either removing or changing unnecessary features or creating a new, smaller set of features (Kumbhar & Mali, 2016). Model training and testing is where machine learning is applied to achieve the desired task (Guo, 2017).

In recent years, the emergence of the transformer model has produced an area of rapid advancement in language modelling (Agarwal, 2019). The transformer model is based on a sequence-to-sequence architecture and implements an attention mechanism within it (Allard, 2019). The attention mechanism recursively determines which words in an input sequence are important to each other, emulating the human thought process of reading. Transformer-based NLP models use this as a feature extraction technique, creating a vector representation for each word in the sequence based on their importance to the other words in the sequence (Ankit, 2020).

Many transformer-based NLP models have surpassed the performance of more traditional machine learning NLP approaches (Wolf et al., 2019). This is largely due to the attention mechanism being used in combination with a neural network deep learning model. Using a large amount of textual data and training

techniques such as masked learning modelling and next sentence prediction, the transformer-based models can build an understanding of language in their neural networks; this is known as *pretraining a model*. These pretrained models can then be fine-tuned by training them on additional data to apply it to a desired application. This opens up the benefits of deep learning to much smaller NLP-type data sets and is a large reason why these models have achieved their high level of performance (Agarwal, 2019).

The field of NLP applications in education, specifically in the assessment of conceptual understanding, is relatively new and has not been widely adopted by educational institutes. There are several studies that have developed differing approaches to the automated assessment of understanding.

In one study, the NLP technique *latent semantic analysis* was used as a similarity comparator between a textual response and idealised peer responses as a means for accurately producing human grading and predicting post-test performance (Guerrero & Wiley, 2019). This approach lacked in its ability to provide insight into a students' conceptions and the nature of their conceptual understanding.

In another study, a combination of NLP techniques and node link representations was used to assess students' understanding in short-response questions (Lajis & Aziz, 2010). This provided a means of performing a similarity comparison at a knowledge level rather than at a textual semantic level. What this entails is that the developed technique has the ability to assess the reproduction of knowledge. As such, a limited insight into the nature of students' conceptual understanding is provided.

In another study, a framework was developed to automatically assesses students' conceptual understanding in adapted concept inventory questions (Cunningham-Nelson, 2019). The questions used were from the signals and systems concept inventory, with a text response field added for students to provide justification for their multiple-choice selection (Wage et al., 2005). NLP techniques were applied to assess whether a student mentioned the correct concept in their response and whether they provided accurate justification. In conjunction with the multiple-choice selection and an algorithm that checked for keywords indicating uncertainty, the model would determine a student's level of conceptual understanding. An accuracy of approximately 85% was achieved with this technique, which is not high enough for reliable use in the classroom.

Research questions

With this background, the following research questions guided this study:

(1) Which NLP techniques can be applied to best extract evidence of conceptual understanding from text?
(2) What performance can be achieved by applying NLP to automatically assess conceptual understanding from students' textual responses?
(3) What impact does the amount of data have on the performance of an automated conceptual understanding assessment model and what implications does this have for future use?

Method

Automated assessment of conceptual understanding approach

Upon investigation of several existing approaches to the automated assessment of understanding, we decided to expand on the techniques developed by Cunningham-Nelson (2019). This is because the approach taken in their study provided a high level of insight into students' conceptual understanding. With the increased language understanding that transformer-based NLP models offer, there was also an opportunity for them to discern additional information from responses that might provide further insight into the nature of students' conceptual understanding. Therefore, we selected transformer-based NLP models for performing the NLP tasks of this study.

The approach developed in this study shares six of the adapted concept inventory questions used in the previous study (Cunningham-Nelson, 2019). We determined that a model would be developed which would assess four pieces of information, called *pointers*, from a student's response and determine a level of conceptual understanding from these. This study commenced with the hypothesis that NLP techniques

could assist in automatically assessing students' conceptual understanding. This was later tested through further investigations that adopted a positivist research paradigm, with statistical measures (namely, accuracy and area under the receiver operating characteristic (ROC) curve) used to evaluate the performance of the model's ability to assess the pointers and hence, conceptual understanding. By using several pointers, a more nuanced level of conceptual understanding in text can be evaluated.

Table 1 details the four pointers and their binary classification classes. The confidence-in-response pointer provides insight into the nature of a student's conceptual understanding: it provides an indication of how strongly formed their conceptions are.

Table 1
Descriptions of the four pointers of conceptual understanding and their binary classification classes

Pointer	Description	Classification classes
Multiple-choice	Whether the student has answered the multiple-choice component correctly	Correct/incorrect
Concept-mentioned	Whether the correct concept or concepts have been mentioned in the student's written justification	Yes/no
Free-text validity	Whether the student's reasoning is valid and correct in their written justification	Correct/incorrect
Confidence-in-response	Level of confidence the student has in their written justification	High/low

With pointer models that can automatically classify each pointer from a response, an overall model would assess the level of a student's conceptual understanding and the level of misconception present on a 5-point scale from *very low* to *very high*. Table 2 displays how the overall classifications are determined from the pointer classifications.

Table 2
How pointer classifications impact the overall level of misconception and conceptual understanding classifications

Multiple-choice	Concept-mentioned	Free-text validity	Confidence-in-response	Level of misconception	Level of conceptual understanding
Incorrect	No	Incorrect	Low	Very low	Very low
Incorrect	No	Incorrect	High	Very high	Very low
Incorrect	No	Correct	Low	Impossible classification	Impossible classification
Incorrect	No	Correct	High	Impossible classification	Impossible classification
Incorrect	Yes	Incorrect	Low	Very low	Low
Incorrect	Yes	Incorrect	High	High	Low
Incorrect	Yes	Correct	Low	Impossible classification	Impossible classification
Incorrect	Yes	Correct	High	Impossible classification	Impossible classification
Correct	No	Incorrect	Low	Very low	Very low
Correct	No	Incorrect	High	Very high	Very low
Correct	No	Correct	Low	Impossible classification	Impossible classification
Correct	No	Correct	High	Impossible classification	Impossible classification
Correct	Yes	Incorrect	Low	Moderate	Moderate
Correct	Yes	Incorrect	High	High	Low
Correct	Yes	Correct	Low	Very low	High
Correct	Yes	Correct	High	Very low	Very high

This dual-output classification model provides valuable insight to educators on both the nature and level of students' conceptual understanding. Areas where students have strongly formed misconceptions (incorrect conceptual understanding with high confidence) can be rapidly identified at very high levels of misconception. This is beneficial to educators as these areas will generally take the greatest effort to address. Similarly, educators can rapidly identify when students have very low levels of conceptual understanding; another situation in which greater effort is likely to be required.

There are two impossible classifications that cannot logically be reached but can occur:

- When the model assesses that the student has provided the correct reasoning in their justification without mentioning the correct concept or concepts in their response. This situation is deemed to be impossible as a correct justification to an adapted concept inventory question requires the student to mention an appropriate concept. This scenario would likely result from a free-text validity misclassification.
- When the model assess that the student has provided the correct reasoning in their justification but selected the incorrect multiple-choice answer. This situation is deemed to be impossible as a student who has provided the correct reasoning would logically have selected the correct multiple-choice answer. This scenario would also likely result from a free-text validity misclassification, rather than student error.

The multiple-choice pointer can be assessed with a simple algorithm which compares the student's selection to the correct answer. The concept-mentioned pointer can also be simply assessed by implementing an algorithm that compares the words in a student's justification to a list of predefined concept words for each adapted concept inventory question.

The free-text validity and confidence-in-response pointers will both require NLP modelling to assess. Due to the simplicity of the multiple-choice and concept-mentioned pointers, this study focused only on the free-text validity and confidence-in-response pointers.

NLP model selection

As there exist a wide variety of transformer-based NLP models suited for different tasks, it was important to select models which are most suitable for the specific application in this study. Within this, it was beneficial to differentiate between suitable models to select those which have the greatest potential for strong performance.

The general language understanding evaluation (GLUE) benchmark (https://gluebenchmark.com/) is a collection of data sets used for training, evaluating and analysing NLP models relative to one another. The public leader board provides an overview of the performance of the ranked models and a human baseline. With the GLUE data sets being varied, it is possible to get an idea of how the ranked models will perform in specific applications. Using the performance of models displayed on the GLUE benchmark, we selected the models best suited and most likely to perform well in classifying the free-text validity and confidence-in-response pointers.

Data collection and preprocessing

The collection of students' responses, to be used for model training and testing, was undertaken in a second-year undergraduate signal analysis course. Students' responses to six adapted concept inventory questions adopted from Cunningham-Nelson's (2019) study were collected between 2015 and 2020. Ethics approval for the collection of student text data was granted by the Queensland University of Technology Human Research Ethics Committee, under approval number 1600000964. The questions were delivered online as non-compulsory coursework to ensure that the ethics considerations were met.

The students' responses to each question were manually classified to create free-text validity and confidence-in-response pointer data sets. We undertook the manual classifications to check for disagreements. Table 3 provides an overview of the number of instances of each class for the question's data sets displaying imbalances and overall sizes. The response lengths ranged from one word to lengthy run-on sentences; the majority of the responses were simple short phrases or sentences.

Table 3
Number of class instances per question for the free-text validity and confidence-in-response data sets

Question	Incorrect justification responses	Correct justification responses	Low confidence responses	High confidence responses
1	95	711	53	753
2	451	168	297	322
3	192	333	110	415
4	383	193	188	388
5	225	130	134	221
6	173	185	80	278

Standard preprocessing techniques were applied to prepare the data for training. To uphold the semantics of students' responses and ensure that the conceptual meaning given from a student remained, minimal preprocessing was done. This is also in line with standards for transformer-based models, where exhaustive preprocessing is unnecessary (Potamias et al., 2020). Lower-casing and punctuation removal was performed on all text. A spell check and autocorrection of misspelled words was then performed using the pyspellchecker Python library (Barrus, 2021). Concept specific, out-of-standard-dictionary words had to be incorporated into the spellcheck dictionary to ensure that these words were not autocorrected incorrectly; examples of such words are Laplace, Fourier, convolution and Nyquist. As the questions were delivered online, there were several cases in each question where duplicate responses were submitted; these likely resulted from students' reattempting the questions to discover the correct multiple-choice answer. As a result, duplicate responses were removed during preprocessing to eliminate any bias which may have occurred as a result of these.

Single word responses were removed from the data sets during preprocessing. This is because it was deemed impossible to justify a multiple-choice selection with a single word. For that reason, a single word response would not provide substantial information to assess confidence or justification accurately. Additionally, many single word responses, such as a string of random characters, were non-valid. These were likely a result of students wanting to check their multiple-choice selection without caring for justifying their response.

As a variety of transformer-based NLP models were tested, the simple transformers library (Rajapakse, n.d.) was used to automatically apply the correct tokeniser for each model. These tokenisers have been specifically designed and chosen to work optimally with each relevant transformer. This was the final stage of preprocessing, converting the response to its required tokens.

Several students provided justification in the form of a question, denoted with a question mark ending their response. This indicated that the student was not confident in their understanding. Therefore, it was decided that for the confidence-in-response pointer modelling, these responses would be automatically classified as non-confident and would be omitted from the model training data sets.

Optimising model training parameters

Machine learning model performance varies substantially with changes to model training parameters. Hence, optimal model parameters for the selected NLP models needed to be found for both the confidence-in-response and free-text validity pointer data sets.

The effects on model performance of adjusting the number of batches and epochs training parameters were assessed. An early stopping algorithm was developed to determine the optimal values for the parameters without overfitting the data. Evaluation loss was used as the early stopping metric with a stopping delta of 0.01 and a stopping patience of 5. The evaluation loss was calculated 5 times in each epoch.

Evaluating performance and ensemble modelling

For the free-text validity pointer model, individual models were trained for each question on their respective data sets. Using a training testing split of 80:20, the accuracy and area under the ROC curve of the selected NLP models with the optimal training parameters were recorded for each question. This was then repeated using smaller subset data sets to investigate the effect of reduced data set sizes. Using these results, the

performance of the different NLP models could be compared. To create a better performing model, the best performing models were then brought together to create an ensemble. This allowed the benefits of each individual model to be leveraged to achieve a higher overall accuracy. It was decided to combine the three best performing models as the ensemble would then provide an overall classification based on majority voting (Rojarath et al., 2016); an overview of an ensemble model is shown in Figure 2. The performance of the ensemble was then compared to the individual models.

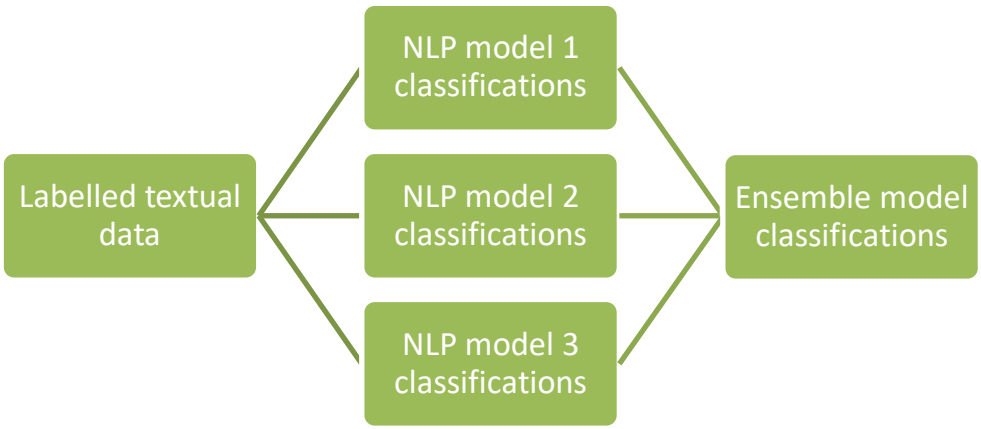

Figure 2. Ensemble NLP model structure

It was decided that confidence in response is not question specific and consequently a non-question-specific model would be developed. The rationale for this is that there were common elements of students' responses across all questions which indicated that they were confident or not confident. To verify this, models would be trained on a data set from one question and then tested against all question's data sets. Using these results, the performance of the different selected NLP models could be compared. An ensemble model would then be created to improve performance.

Results

Investigating GLUE benchmark leader board

The free-text validity and confidence-in-response pointers require the evaluation of the meaning of a sentence or phase, so a strong understanding of language is required. Upon evaluation of the GLUE benchmark data sets, it was decided that all data sets except the corpus of linguistic acceptability required classification based on text meaning. Hence, the performance of the ranked NLP models on all data sets except the linguistic acceptability data set were used to assess which would likely perform well on classifying free-text validity and confidence-in-response. Seven sequence classification NLP models available in the open-source Hugging Face Transformers Library (Hugging Face, 2020) are ranked on the GLUE benchmark leader board:

- ELECTRA
- RoBERTa
- BERT
- MobileBERT
- XLNet
- ALBERT
- XLM.

The ELECTRA, RoBERTa, XLNET and ALBERT models have a greater average accuracy than the human baseline on the GLUE data sets requiring classification based on text meaning. Due to the relatively small size of the data sets available in this study, the smallest versions of these models were selected to model free-text validity and confidence-in-response: ELECTRA-small, RoBERTa-base, XLNet-base and ALBERT-base-v2.

Modelling free-text validity

To evaluate the impact of the number of batches and epochs training parameters, a range of values was chosen and tested for the selected models on each question's free-text validity data set. The following parameter values were tested:

- The number of epochs was incremented from 1 to 10 in steps of 1, while the number of batches was kept at 8.
- The number of batches was incremented from 1 to 200 in steps of 50, with the number of epochs set to 1, 5 and 10.

It was found that the performance of the models varied with the number of epochs but was independent of the number of batches. With the number of batches having no impact on model performance, it was kept constant at its default value of 8 for all modelling. The number of batches parameter had no impact due to the relatively small training data sets used throughout. With larger data sets, the effect of adjusting the number of batches parameter may become noticeable.

An early stopping algorithm was designed to determine the optimal number of epochs for each of the selected NLP models. The results of early stopping are displayed in Table 4.

Table 4
Optimal number of epochs identified by the early stopping algorithm for each model and question free-text validity data set

Model	No more improvement after epoch number					
	Question number					
	1	2	3	4	5	6
ELECTRA-small	1	3	6	3	5	6
RoBERTa-base	1	3	3	3	4	5
XLNet-base	1	2	3	3	5	4
ALBERT-base-v2	1	3	3	2	3	4

To investigate how the optimal number of epochs varied with a reduction in the amount of available data, subset data sets were created for each question by taking a random sample of 100 correct and 100 incorrect justification responses. An equal number of responses from each class was selected to ensure that the training data was balanced and to avoid overfitting. The early stopping results for the subset data sets are displayed in Table 5.

Table 5
Optimal number of epochs identified by the early stopping algorithm for each model and question free-text validity subset data set

Model	No more improvement after epoch number					
	Question number					
	1	2	3	4	5	6
ELECTRA-small	6	5	5	6	6	3
RoBERTa-base	4	3	4	5	3	1
XLNet-base	4	3	4	1	4	1
ALBERT-base-v2	3	4	3	3	3	2

Comparing the results displayed in Tables 4 and 5, there are slight differences in the optimal number of epochs. To determine an optimal number of epochs which is independent of data set size, a sample of values was selected for each model based on the results presented in Tables 4 and 5.

For the ELECTRA-small model, six epochs was selected as the optimal parameter based on its consistent appearance in the early stopping results. The performance of the ELECTRA-small model trained with six epochs is displayed in Table 6.

Table 6
ELECTRA-small model performance with six epochs, trained on each question's free-text validity data set and subset

Data set	Accuracy (%)	Area under ROC curve
Question 1	93.96	0.6630
Question 2	93.67	0.9465
Question 3	92.31	0.8762
Question 4	**94.29**	**0.9466**
Question 5	90.24	0.9055
Question 6	90.57	0.8973
Data set average	**92.51**	**0.8725**
Question 1 subset	85.71	0.8937
Question 2 subset	84.85	0.8212
Question 3 subset	92.31	0.8571
Question 4 subset	90.00	0.8937
Question 5 subset	91.67	0.9063
Question 6 subset	**92.31**	**0.9226**
Subset average	**89.48**	**0.8824**

Note. The best results in model performance are highlighted in bold.

For the RoBERTa-base model, three and four epochs were selected for testing based on the early stopping results. The performance of the RoBERTa-base model trained with three and four epochs is displayed in Table 7.

Table 7
RoBERTa-base model performance with three and four epochs, trained on each question's free-text validity data set and subset

Data set	3 epoch accuracy (%)	3 epoch area under ROC curve	4 epoch accuracy (%)	4 epoch area under ROC curve
Question 1	93.96	0.6630	93.96	0.7010
Question 2	91.14	0.9099	92.41	0.9204
Question 3	92.31	0.8254	91.03	0.7921
Question 4	**92.86**	**0.9350**	**95.71**	**0.9651**
Question 5	92.68	0.9282	95.12	0.9510
Question 6	86.79	0.8497	90.57	0.8891
Data set average	**91.62**	**0.8519**	**92.97**	**0.9035**
Question 1 subset	89.29	0.8918	82.14	0.8392
Question 2 subset	84.85	0.8346	87.88	0.8596
Question 3 subset	88.46	0.7857	96.15	0.9737
Question 4 subset	**93.33**	**0.9231**	**96.67**	**0.9615**
Question 5 subset	87.50	0.8438	87.50	0.8438
Question 6 subset	84.62	0.8333	92.31	0.9167
Subset average	**88.01**	**0.8521**	**90.44**	**0.8991**

Note. The best results in model performance are highlighted in bold.

For the XLNet-base model, three and four epochs were selected for testing based on the early stopping results. The performance of the XLNet-base models trained with three and four epochs is displayed in Table 8.

Table 8
XLNet-base model performance with three and four epochs, trained on each question's free-text validity data set and subset

Data set	3 epoch accuracy (%)	3 epoch area under ROC curve	4 epoch accuracy (%)	4 epoch area under ROC curve
Question 1	93.96	0.6602	93.96	0.6630
Question 2	93.67	0.9308	93.67	0.9422
Question 3	89.74	0.7333	91.03	0.8175
Question 4	98.57	0.9884	97.14	0.9767
Question 5	**100.00**	**1.000**	**97.56**	**0.9737**
Question 6	83.02	0.8185	83.02	0.8348
Data set average	**93.16**	**0.8552**	**92.73**	**0.8680**
Question 1 subset	85.71	0.8363	82.14	0.7807
Question 2 subset	87.88	0.8462	90.91	0.8846
Question 3 subset	88.46	0.8308	88.46	0.8308
Question 4 subset	**90.00**	**0.8846**	**96.67**	**0.9615**
Question 5 subset	84.62	0.8512	87.50	0.8438
Question 6 subset	84.62	0.8512	96.15	0.9583
Subset average	**86.88**	**0.8501**	**90.31**	**0.8766**

Note. The best results in model performance are highlighted in bold.

For the ALBERT-base-v2 model, three and four epochs were selected for testing based on the early stopping results. The performance of the ALBERT-base-v2 models trained with three and four epochs is displayed in Table 9.

Table 9
ALBERT-base-v2 model performance with three and four epochs, trained on each question's free-text validity data set and subset

Data set	3 epoch accuracy (%)	3 epoch area under ROC curve	4 epoch accuracy (%)	4 epoch area under ROC curve
Question 1	93.96	0.663	91.95	0.576
Question 2	94.94	0.9583	94.94	0.9583
Question 3	85.9	0.6841	88.46	0.7762
Question 4	**95.71**	**0.9651**	**95.71**	**0.9651**
Question 5	87.8	0.8792	90.24	0.9055
Question 6	73.58	0.7076	96.05	0.9643
Data set average	**88.65**	**0.8096**	**92.89**	**0.8576**
Question 1 subset	71.43	0.7018	78.57	0.7544
Question 2 subset	87.88	0.8731	87.88	0.8462
Question 3 subset	73.08	0.6353	80.77	0.688
Question 4 subset	90	0.8937	93.33	0.9231
Question 5 subset	79.17	0.8125	83.33	0.8438
Question 6 subset	**92.31**	**0.9167**	**96.15**	**0.9643**
Subset average	**82.31**	**0.8055**	**86.67**	**0.8366**

Note. The best results in model performance are highlighted in bold.

Based on the performance results presented in Tables 6, 7, 8 and 9, the ELECTRA-small model trained with six epochs, RoBERTa-base model trained with four epochs and the XLNet-base model trained with four epochs produced the strongest results. Therefore, these three models were selected for an ensemble. On a computer with a 6-core 12-thread central processing unit (CPU) and 128 gigabytes (GB) of random-access memory (RAM), it took an average of 39 minutes to train the free-text validity ensembles, ranging from 57 minutes for the largest data set to 27 minutes for the smallest. It should be noted that the ensembles could not be trained on a computer with 8 GB of RAM due to insufficient memory. The performance of the ensemble model on each question's data set is displayed in Table 10.

Table 10
Free-text validity ensemble model performance on question data sets using 5-fold cross-validation

Data set	Accuracy (%)	Area under ROC curve
Question 1	97.97	0.8641
Question 2	97.34	0.9747
Question 3	**98.66**	**0.9814**
Question 4	96.86	0.9661
Question 5	93.54	0.9376
Question 6	95.95	0.9577
Average	**96.72**	**0.9469**

Note. The best results in model performance are highlighted in bold.

The performance of the ensemble exceeded the performance of the individual models. The average accuracy of 96.72% and area under the ROC curve of 0.9469 indicate that the model could distinguish between correct and incorrect justification well and accurately. With a study indicating that educators assess their students' understanding with accuracies lower than this (Chi et al., 2004), the ensemble model provides very promising results.

To evaluate the ensemble's suitability and adaptability to smaller data sets, its performance was also found on two subset data sets for each question: one subset comprised of 100 correct and incorrect justification responses and another with 40 correct and incorrect justification responses. The performance of the ensemble model on these data sets is displayed in Tables 11 and 12 respectively.

Table 11
Ensemble model performance on subsets of 100 correct and incorrect justification data sets

Data set	Accuracy (%)	Area under ROC curve
Question 1 subset	91.46	0.8976
Question 2 subset	**97.95**	**0.9770**
Question 3 subset	97.78	0.9707
Question 4 subset	97.62	0.9756
Question 5 subset	96.35	0.9583
Question 6 subset	97.09	0.9691
Average	**96.38**	**0.9581**

Note. The best results in model performance are highlighted in bold.

Table 12
Ensemble model performance on subsets of 40 correct and incorrect justification data sets

Data set	Accuracy (%)	Area under ROC curve
Question 1 subset	92.47	0.8985
Question 2 subset	**98.33**	**0.9500**
Question 3 subset	93.57	0.8607
Question 4 subset	96.00	0.9524
Question 5 subset	94.00	0.8875
Question 6 subset	94.07	0.9578
Average	**94.74**	**0.9178**

Note. The best results in model performance are highlighted in bold.

The performance of the ensemble model was found to be very strong with high accuracies and area under the ROC curves being achieved across all subset data sets. Effectively identical performance was achieved between the ensembles trained on the complete data sets and the subsets of 100 correct and incorrect justification responses, with only a drop in average accuracy of 0.34%. Furthermore, there was a minimal drop of 1.98% in accuracy on the significantly smaller subset of 40 correct and incorrect justification responses. This indicates that educators wishing to adapt this automated conceptual understanding assessment approach into their own classes can do so with minimal data and achieve strong, human-like performance.

Table 10 shows that the free-text validity ensemble trained on the entire question data sets had the weakest performance on the last three questions. This can be explained by the relative sizes of the data sets for each question. The first three data sets contained more responses than the last three. As these models gain an understanding of how to classify data through training, they improve at classifying with more responses. Therefore, the models trained on large data sets (the first three questions) have stronger performance than those trained on smaller data sets (the last three question). This is reflected in Tables 11 and 12; when the models are trained on equal numbers of student responses, the performance is similar across the questions.

Modelling confidence-in-response

Confidence is another key pointer which provides valuable insight into the nature of students' conceptual understanding. The number of batches parameter was kept constant at 8 as it had not impacted model performance during the free-text validity modelling. The early stopping algorithm developed in the free-text validity modelling was used to determine the optimal number of epochs for the confidence models. The question 1 data set was arbitrarily selected to develop the confidence model. Due to the large imbalance in favour of the high confidence responses in the Question 1 data set, it was assumed that a substantial bias would exist without rebalancing measures. As such, a training data set was created, consisting of a random sample 50 high confidence responses and all 53 low confidence responses. The early stopping results of the models on the training data set are displayed in Table 13.

Table 13
Optimal number of epochs identified by the early stopping algorithm for each model on the Question 1 training data set

Model	No more improvements after epoch number
ELECTRA-small	7
RoBERTa-base	4
XLNet-base	5
ALBERT-base-v2	5

Each question's confidence in response data set was combined to form a large testing data set used to evaluate the performance of the confidence in response models. With the models trained on the training data set with the optimal number of epochs, the performance of each on the testing data set was found; the results are displayed in Table 14.

Table 14
Performance of optimal epoch models on the testing data set when trained on the Question 1 training data set

Model	Accuracy (%)	Area under ROC curve
ELECTRA-small	87.89	0.8738
RoBERTa-base	**91.24**	**0.8663**
XLNet-base	88.13	0.8983
ALBERT-base-v2	86.50	0.8897

Note. The best results in model performance are highlighted in bold.

Based on the performance results presented in Table 14, the RoBERTa-base model trained with four epochs, the XLNet-base model trained with five epochs and the ALBERT-base-v2 model trained with five epochs produced the strongest results. As a result, these three models were selected for an ensemble. Although the accuracy of the ELECTRA-small model exceeded that of the ALBERT-base-v2 model slightly, the ALBERT-base-v2 model had the second highest area under ROC curve and was hence preferentially selected. On a computer with a 6-core 12-thread CPU and 128 GB of RAM, it took approximately 13 minutes to train this confidence-in-response ensemble. It should be noted that the ensemble could not be trained on a computer with 8 GB of RAM due to insufficient memory. The performance of the ensemble model on each question's data set is displayed in Table 15.

Table 15
Performance of the confidence-in-response ensemble on each question's data set when trained on the Question 1 training data set

Data set	Accuracy (%)	Area under ROC curve
Question 1	**96.72**	**0.9834**
Question 2	91.87	0.9059
Question 3	91.96	0.9576
Question 4	80.89	0.8946
Question 5	79.46	0.8644
Question 6	91.71	0.8780
Average	**88.77**	**0.9140**

Note. The best results in model performance are highlighted in bold.

The ensemble model produced an average accuracy of 88.77% which is only slightly above the average accuracy of the XLNet-base and ALBERT-base-v2 models and below the accuracy of the RoBERTa-base model. Table 15 shows that the performance of the ensemble is very strong on all questions except 4 and 5. When inspecting the misclassification in these two questions, it was clear that many contained content-specific, non-standard-dictionary words. A potential reason for the drop in performance could be the nature of the selected NLP models; they rely heavily on their understanding of language from pretraining. It seems feasible to expect that despite being pretrained on large data sets, the relative frequency of the content-specific words would be very small in comparison to standard-dictionary words; so, it could be expected that the pretrained models have a limited understanding of these content-specific words. When investigating the responses in the other questions, they rarely featured these content-specific words. With the ensemble being trained on responses that do not feature the content-specific words, there is no way for the ensemble to build an understanding of the words and hence it would struggle classifying them.

To test this hypothesis and strive for stronger performance with another question from the original data set, we tested the model with a training data set from the Question 5 responses. The training data set consisted of a random sample of 50 high confidence responses and all low confidence responses (35 after preprocessing). This ensemble took 10 minutes to train on a computer with a 6-core 12-thread CPU and 128 GB of RAM. The performance of the ensemble trained on this data set on each question's data set is displayed in Table 16.

Table 16
Performance of the confidence-in-response ensemble on each question's data set when trained on the Question 5 training data set

Data set	Accuracy (%)	Area under ROC curve
Question 1	98.52	0.6841
Question 2	93.07	0.8387
Question 3	98.07	0.9012
Question 4	96.75	0.8456
Question 5	**99.46**	**0.9800**
Question 6	97.41	0.7917
Average	**97.21**	**0.8402**

Note. The best results in model performance are highlighted in bold.

The performance of this ensemble was significantly greater than the first. With an average accuracy of 97.21% and area under the ROC curve of 0.8402, the ensemble could distinguish between high and low confidence responses accurately and well. With this performance exceeding that of the free-text validity ensembles, it indicates that free-text validity is more complex and challenging to model. Furthermore, as explained in the free-text validity results, these accuracies are greater than the accuracy of which educators assess their students' understanding (Chi et al., 2004), suggesting that the automated approach developed in this study may provide more consistent and accurate assessments.

The ensemble's deficiencies with the Questions 4 and 5 data sets were eliminated by training the model on the Question 5 training data set, with significant jumps in performance. These results suggest that the pretrained NLP models may lack an understanding of content-specific, non-standard-dictionary words; however, this can be accounted for during training.

The strong performance of the ensemble being achieved with a small training data set indicates that in future applications, a confidence-in-response ensemble model can be developed without its performance being restricted to large data sets.

Discussion

The results of the free-text validity and confidence-in-response modelling indicate that the tested NLP modelling techniques can be applied to accurately assess these two pointers of students' conceptual understanding in their short answer textual responses. This is due to the strong performance of the ensemble pointer models, exceeding an average accuracy of 95% in combination with high area under the ROC curves. In conjunction with simple algorithms which determine the other two pointers, students' multiple-choice selection and concept-mentioned in justification, the automated assessment of conceptual understanding technique presented can be realised and implemented with accuracies exceeding 95%.

As explained previously, accuracies above 95% are at least on par with educator assessments of conceptual understanding (Chi et al., 2004). Hence, the techniques developed in this study can be applied by educators as a formative assessment approach which can perform to a human standard. Additionally, with the approach being completely automated, the techniques can be readily applied into flexible and online educational environments, presenting a new formative assessment option for educators with the added benefits of fast feedback as and when required for themselves and students.

Of the automated assessment approaches that have been reviewed, the approach developed in this study provides a unique insight into students' conceptual understanding with more accurate performance. The incorporation of a justification text field allows for insight into the nature of students' conceptual understanding beyond traditional multiple-choice concept inventory questions. Furthermore, the approach which this study expanded on assessed students' conceptual understanding with accuracies around 85% (Cunningham-Nelson, 2019); our approach exceeds this performance considerably in all pointers. By designing our approach to assess conceptual understanding from pointers, rather than similarity to an exemplar response, it also provides substantially greater insight when compared to other automated understanding assessment approaches (Guerrero & Wiley, 2019; Lajis & Aziz, 2010).

The performance of the free-text validity pointer ensembles experienced minimal loss when the significantly smaller subset data sets were used. The confidence-in-response ensembles achieved strong performance despite the very small training data sets used. These observations indicate that educators wishing to adapt this automated conceptual understanding assessment approach into their own classes can do so with minimal data and achieve strong, human-like performance.

Training the ensemble models took an average of 39 minutes for the free-text validity model and 11.5 minutes for the confidence-in-response model. By using pretrained models, the training (fine-tuning) time is fast when compared to training an entire transformer model and also allows for the utilisation of the language understanding that exists from their pretraining. The training times related to the size of the training data sets, with smaller data sets training significantly faster than the larger data sets. As training was done on a relatively powerful computer which had 128 GB of RAM, training would take longer on less powerful machines. Furthermore, machines with relatively low memory may be incapable of training such models; a computer with 8 GB of RAM failed to train all ensembles. This is due to the complexity and size of the transformer models as they require substantial computational resources. Fortunately, all ensemble models can also be trained on various free-to-use online platforms. This means that educators have the tools to train their own models, regardless of the computational resources at their disposal.

With the smallest data set size having been tested containing 80 responses, educators should note that performance on smaller data sets is unknown and likely to decrease. Importantly, educators should be wary when responses are expected to contain content-specific words as the performance of the confidence-in-response models was limited in cases where the training data did not contain such words. Therefore, to

achieve strong performance, responses containing content-specific words should be incorporated into the training data set.

In applications where greater performance is required, educators could look at utilising larger data sets, which would lead to stronger performance. It should be noted that data sets containing a large class imbalance are expected to create bias and hence weaker performance. Therefore, if a large imbalance is present, we recommend that educators sample out a balanced data set for model training purposes. Additionally, greater performance may be achieved through further optimisation of training parameters as only the number of batches and epochs were optimised in this study.

There were some common responses that the free-text validity and confidence-in-response models struggled with. Educators should be aware that both models would sometimes misclassify responses containing multiple non-valid words or phrases. An algorithm could be developed to deal with them during preprocessing.

There were also other limitations which should be taken into consideration for future applications. The data sets all came from a single subject. Although there are differences in concepts between questions, they share similarities within the subject area of signal analysis. The models would perform strongly in other subject areas; however, this has not been validated. Additionally, further research needs to be conducted to assess how the models perform with more complex multi-sentence responses.

Conclusion

This study developed a technique to automatically assess students' conceptual understanding from their responses to short response questions. Specifically, modelling the free-text validity of a student's justification as well as the confidence in their response have been addressed. By training several open-source NLP models, question specific ensemble models that could assess the validity of a student's justification to accuracies exceeding 90% could be created. Furthermore, the detrimental effects on model performance due to a reduction in the amount of available training data were minimal. A confidence-in-response model was developed which achieved accuracies above 95% on all question data sets.

Transformer-based NLP models can be used to best extract evidence of conceptual understanding from text. By combining several of these models, human-like performance for the assessment of conceptual understanding can be achieved. The amount of available data had minimal impact on the performance of the models, suggesting that human-like conceptual understanding assessment models can be developed by educators with access to small numbers of student responses. This study complements the existing body of work of applied NLP in education, showcasing a novel approach to assessing conceptual understanding.

The insights gained from this study indicate a promising future for NLP applications in education. With the viability of the automated assessment of conceptual understanding realised in this study, the advantages of this approach to formative assessment may soon be realised and adopted widely, providing opportunities for enhanced learning experiences for both educators and students alike.

Acknowledgements

Computational resources and services used in this work were provided by HPC and the Research Support Group, Queensland University of Technology, Brisbane, Australia.

References

Agarwal, N. (2019, July 2). Examining the transformer architecture – Part 2: A brief description of how transformers work. *KDnuggets*. https://www.kdnuggets.com/2019/07/transformer-architecture-part-2.html

Alammar, J. (2019, March 27). The illustrated word2vec. *GitHub*. https://jalammar.github.io/illustrated-word2vec/

Allard, M. (2019, January 25). What is a transformer? — Inside machine learning. *DZone*. https://dzone.com/articles/what-is-a-transformer-inside-machine-learning

Ankit, U. (2020, April 25). Transformer neural network: Step-by-step breakdown of the beast. *Towards Data Science*. https://towardsdatascience.com/transformer-neural-network-step-by-step-breakdown-of-the-beast-b3e096dc857f

Barrus, T. (2021). *Pyspellchecker 0.6.2*. Python Software Foundation. https://pypi.org/project/pyspellchecker/

Burstein, J. (2009). Opportunities for natural language processing research in education. In A. Gelbukh (Ed.), *Lecture notes in computer science: Vol. 5449. Computational linguistics and intelligent text processing* (pp. 6–27). Springer. https://doi.org/10.1007/978-3-642-00382-0_2

Chi, M., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363–387. https://doi.org/10.1207/s1532690xci2203_4

Cunningham-Nelson, S. (2019). *Enhancing student conceptual understanding and learning experience through automated textual analysis* [Doctoral dissertation, Queensland University of Technology]. ePrints. https://doi.org/10.5204/thesis.eprints.134145

D'Avanzo, C. (2008). Biology concept inventories: Overview, status, and next steps. *BioScience*, *58*(11), 1079–1085. https://doi.org/10.1641/b581111

Denny, J. C., Bastarache, L., Sastre, E. A., & Spickard, A. (2009). Tracking medical students' clinical experiences using natural language processing. *Journal of Biomedical Informatics*, *42*(5), 781–789. https://doi.org/10.1016/j.jbi.2009.02.004

Dodge, J. (n.d.). *What are formative assessments and why should we use them?* Scholastic. https://www.scholastic.com/teachers/articles/teaching-content/what-are-formative-assessments-and-why-should-we-use-them/

Ganesan, K. (2019, April 9). All you need to know about text preprocessing for NLP and machine learning. *KDnuggets*. https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

Garbade, M. J. (2018, October 15). A simple introduction to natural language processing. *Becoming Human*. https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

Gikandi, J., Morrow, D., & Davis, N. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Goncher, A. M., & Boles, W. (2017). Enhancing the effectiveness of concept inventories using textual analysis: Investigations in an electrical engineering subject. *European Journal of Engineering Education*, *44*(1-2), 222–233. https://doi.org/10.1080/03043797.2017.1410523

Guerrero, T., & Wiley, J. (2019). Using "idealized peers" for automated evaluation of student understanding in an introductory psychology course. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Lecture notes in computer science: Vol. 11625. Artificial intelligence in education* (pp. 133–143). Springer. https://doi.org/10.1007/978-3-030-23204-7_12

Guo, Y. (2017, September 7). The 7 steps of machine learning. *Towards Data Science*. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hugging Face. (2020). *Transformers*. Retrieved November 18, 2020, from https://huggingface.co/transformers/

Keeley, P. (2008). *Science formative assessment* (1st ed.). Corwin.

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, *20*(2), 177–185. https://doi.org/10.1016/j.knosys.2006.11.014

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), Article 150. https://doi.org/10.3390/info10040150

Krithika, R., & Narayanan, J. (2015). Learning to grade short answers using machine learning techniques. In I. Nair (Ed.), *Proceedings of the Third International Symposium on Women in Computing and Informatics* (pp. 262–271). Association for Computing Machinery. https://doi.org/10.1145/2791405.2791508

Kumbhar, P., & Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research*, *5*(5), 1267–1275. https://doi.org/10.21275/v5i5.nov163675

Lajis, A., & Aziz, N. A. (2010). NL scoring technique for the assessment of learners' understanding. In *Proceedings of the Second International Conference on Computer Research and Development* (pp. 379–383). IEEE. https://doi.org/10.1109/iccrd.2010.68

List, A., & Alexander, P. (2015). Examining response confidence in multiple text tasks. *Metacognition and Learning*, *10*(3), 407–436. https://doi.org/10.1007/s11409-015-9138-2

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Best practices for administering concept inventories. *The Physics Teacher*, *55*(9), 530–536. https://doi.org/10.1119/1.5011826

McCarthy, J. (2017). Enhancing feedback in higher education: Students' attitudes towards online and in-class formative assessment feedback models. *Active Learning in Higher Education*, *18*(2), 127–141. https://doi.org/10.1177/1469787417707615

Potamias, R. A., Siolas, G., & Stafylopatis, A. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

Rajapakse, T. (n.d.). *Simple transformers*. https://simpletransformers.ai/

Rojarath, A., Songpan, W., & Pong-inwong, C. (2016). Improved ensemble learning for classification techniques based on majority voting. In *Proceedings of the 7th IEEE International Conference on Software Engineering and Service Science* (pp. 107–110). IEEE. https://doi.org/10.1109/icsess.2016.7883026

Smith, G. G., Haworth, R., & Žitnik, S. (2020). Computer science meets education: Natural language processing for automatic grading of open-ended questions in ebooks. *Journal of Educational Computing Research*, *58*(7), 1227–1255. https://doi.org/10.1177/0735633120927486

Sultan, M., Salazar, C., & Sumner, T. (2016). Fast and easy short answer grading with high accuracy. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1070–1075). Association for Computational Linguistics. https://doi.org/10.18653/v1/n16-1123

Survey Anyplace. (n.d.). *Multiple-choice question*. https://help.surveyanyplace.com/en/support/solutions/articles/35000042297-multiple-choice-question

Wage, K. E., Buck, J. R., Wright, C. H. G., & Welch, T. B. (2005). The signals and systems concept inventory. *IEEE Transactions on Education*, *48*(3), 448–461. https://doi.org/10.1109/te.2005.849746

Widmann, M., & Silipo, R. (2015, May 12). Seven techniques for data dimensionality reduction. *KNIME*. https://www.knime.com/blog/seven-techniques-for-data-dimensionality-reduction

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Association for Supervision and Curriculum Development.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019). *HuggingFace's transformers: State-of-the-art natural language processing*. Retrieved October 12, 2020, from https://arxiv.org/abs/1910.03771

Corresponding author: Rick Somers, rsomers122@outlook.com

Please cite as: Somers, R., Cunningham-Nelson, S., & Boles, W. (2021). Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australasian Journal of Educational Technology*, *37*(5), 98-115. https://doi.org/10.14742/ajet.7121