Natural Language Processing Skills of Learnings Through the Use of Conceptual Interpretation and Textual Responses

Ramlan Mustapha[1], Sti Habsah Sh Zahari[2], Ahmad Zam Hariro[3], A. Jailani Che Abas[4], Norsafizar Mohd Noor[5], Nor Adila Mohd Noor[6], Senin MS[7], Zaharuddin Ibrahim[8]

[1]Universiti Teknologi MARA Pahang, Raub Campus, Malaysia
[2,3,8]Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[4]Institut Pendidikan Guru Kampus Tengku Ampuan Afzan, Pahang, Malaysia
[5]Xiamen University Malaysia
[6]Academy of Contemporary Islamic Studies, Universiti Teknologi MARA Terengganu, Dungun Campus, Malaysia
[7]ndependent Researcher, Malaysia

Abstract:

The goal of this study was to see how natural language processing (NLP) approaches could be used in an educational setting to assess Learners' conceptual knowledge based on their short answer responses. Completion testing stimulates and offers response on Learners' abstract knowledge, which is frequently overlooked in automated grading. Automated formative assessment, which provides insights into conceptual comprehension as needed, benefits both Learners and instructors, especially in online education and large cohorts. It employed the ELECTRO-small, Roberto-base, XXLNET-base, and ALBERTO-V3 NLP machine learning models. These two parts of data shed light on Learners' conceptual understanding as well as the nature of their comprehension. It used high-performance NLP models to build a free-text validity ensemble with accuracies ranging from 91.46 percent to 98.66 percent for judging the validity of Learners' justifications. With precisions ranging from 93.07 percent to 99.46 percent, it suggested a generic, non-question-specific Response model for categorizing responses as high or low confidence. Because of the great presentation of these models and their adaptability to lesser data sets, instructors have an exclusive chance to incorporate these approaches into their lectures.

Contextual or rule implications:

- Learners' conceptual comprehension can be precisely and automatically determined by analyzing their short answer responses using natural language processing.
- Instructors and Learners can receive feedback on conceptual comprehension on a need-to-know basis via automated assessment, without incurring the overhead associated with traditional formative evaluation.

Instructor can use automated assessment to accurately evaluate conceptual comprehension models when they receive less than 100 responses to their short answer questions from Learners.

Keywords: NLP, automated comprehension assessment, formative assessment, machine learning, conceptual comprehension, and mixed approaches are some of the terms that come to mind while thinking about NLP

1. INTRODUCTION

Instructor can use automated assessment to accurately evaluate conceptual comprehension models when they receive less than 100 responses to their short response questions from Learners.

©2021JPPW.Allrights reserved

Traditional formative evaluation approaches are becoming less useful and acceptable in today's schools as flexible study arrangements become more widespread (Wang et al., 2020). Computerized assessment enables instructors and Learners to get feedback consistently as and when it is necessary. Additionally, a computerized method mitigates the negative consequences of growing class numbers by considerably reducing the time needed for conventional evaluation.

There are numerous educational applications for natural language processing, including dialogue-based instruction, paraphrasing tools, and text quality software (Goularte et al.,2019). NLP applications include educational chatbots (Fernando, 2020), (Tapingkae et al., 2020) recommend using automated grading systems and technologies to keep track of educational experiences (Granberg et al. 2021). These few NLP samples were used to determine a greater level of comprehension. This learning sought to supplement previously published research on the use of NLP in education by examining its potential for autonomously evaluating Learners' conceptual knowledge.

Historically, automated assessment systems have placed a premium on Learners' work being properly marked and graded. Automated understanding evaluation is unique in that it gives Learners precise feedback on their conceptual understanding and allows them to self-assess and revisit their knowledge whenever they want, independent of instructor availability or time constraints. Additionally, Learner’s benefit from the opportunity to evaluate conceptual knowledge; they can quickly identify and correct errors, as well as check their understanding of concepts, allowing them to develop confidence in their comprehension.

In minimizing the amount of time spent on conventional formative assessment, instructors get additional advantages by incorporating automated evaluation of Learners' conceptual knowledge into their instruction. Instructors may be able to devote more time to teaching, improving instructional tactics and resources, and responding to student questions and concerns. Instructors might alter their course based on input about their cohort's conceptual knowledge to increase conceptual comprehension most effectively while correcting misconceptions. Additionally, this feedback provides Instructor with an opportunity to reflect on and improve their current and future teaching strategies and materials.

Learners benefit from formative evaluation because it gives vital feedback on their conceptual knowledge. With more significant cohorts, instructors have limited time to evaluate Learners effectively and provide relevant feedback (Broadbent et al., 2021). Additionally, since online education reduces face-to-face time, conventional formative evaluation, which gives Learners rapid feedback on their conceptual knowledge, diminishes. Natural language processing enables the extraction of critical insights into pupils' conceptual comprehension, enabling an automated evaluation technique.

Literature review

Evaluation of conceptual comprehension

One way to evaluate a person's conceptual grasp is to examine their skill to apply their knowledge and talents in unexpected settings and situations (Chen et al. 2020). This proof may be gathered via proper classroom evaluation. Assessments of conceptual comprehension must be constructed so that evidence of transferability may be recognized. As a result, the assessment's design is critical to its capacity to provide this evidence.

Formative assessment is beneficial because it enables instructors and Learners to get feedback that guides their choices toward achieving learning objectives. Using formative evaluation, Instructor may ascertain Learners' abilities, knowledge, and conceptual comprehension. By imparting these insights to pupils, it enables them to direct their own self-education. The feedback pupils get is very beneficial since individuals are often inaccurate in their assessment of what they do and do not know (Schildkamp et al., 2020).

Teachers may improve learning outcomes by employing a constructivist teaching strategy with proof of Learners' abilities, knowledge, and conceptual understanding (Zainuddin et al., 2020). By assessing assumptions about certain subjects, to successfully establish and develop them, learning can be focused. Evaluating Learners' current beliefs can reveal the correctness of their knowledge, talents, and conceptual understanding, as well as the existence or absence of misunderstandings.

Automated evaluation techniques

Educational institutions have created and utilized various automated formative and summative evaluation approaches. However, most cannot measure conceptual comprehension due to their nature; those that were frequently erroneous or lacked the ability to reveal the quality of pupils' conceptual knowledge.

Questions tests are popular because they are objective and can be easily automated to provide learners and teachers with instant feedback. Special attention must be made to the questions and accessible responses in order to get over the assessment's inherent conceptual comprehension limitations. Using Questions, concept inventories are used to assess learners' understanding of certain concepts (Veugen et al. 2021). By combining practical questions with appropriately prepared distractors, these assessments can assess Learners' comprehension and reveal misunderstandings. Despite this, they are incapable of evaluating guessed choices, establishing the source or logic of Learners' misunderstandings, or providing insight into the nature of Learners' comprehension (Liu Et al. (2021). Applied natural language processing in teaching, particularly in the assessment of conceptual comprehension, is still in its early phases, with only a few educational institutions embracing it.

NLP is used to extract meaning from texts.

The difficulty with NLP, particularly in the education arena, is that the original text's semantic meaning must be preserved in its numerical representation. Since computers cannot directly comprehend language, they cannot extrapolate meaning from it (Zhu et al., (2020). Thus, the goal of NLP is to convert text to an interpretable arithmetic illustration. The general stages of NLP deployment in the real world are depicted in the NLP process diagram (see Figure 1).

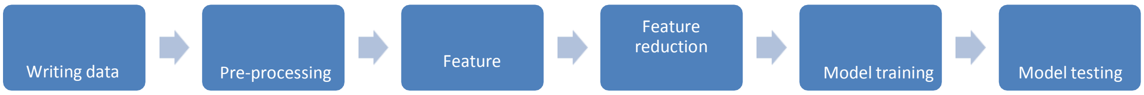


Figure 1. The assembly of the Natural Language Processing (NLP) process (adapted from El-Kassas et al. 2021)

Preprocessing aims to enhance the text's predictability and analyzability (Ray et al. 2020). In the natural language processing systems, lower-casing and punctuation removal are both necessary pre-processing processes. Several approaches, depending on the presentation and total amount of text data available, techniques such as reducing, lemmatization, and text enrichment may be useful. By using feature extraction, the preprocessed text is converted to numerical data (Martínez & Ruiz (2020). Text characteristics might be as simple as the text's word count or as complicated as vector representations of the words (Schellekens et al. 2021). Feature reduction may be advantageous depending on the approach used to extract the features. The purpose of feature reduction is to compact numerical data to make it more interpretable (Ma & Xu, (2020). Generally, feature reduction strategies exclude or alter superfluous characteristics or establish a new, more compact collection of features (Pallathadka et al., 2021). Model training and validation are the processes using machine learning to accomplish a goal (Rahman et al. (2021).

The introduction of the transformer model has accelerated the progress of language modeling (Zeineddine et al., 2021). The transformer model has a mechanism for paying attention and is built on a sequence-to-sequence architecture (Vartiainen et al. (2021). The attention mechanism detects whether words in an input sequence are related recursively, simulating the human reading process. This method for feature extraction leveraging the relationship between words to build a vector presentation for each word in the arrangement is used by transformer-based natural language processing models (Wu) (2021).

Numerous transformer-based natural language processing models have outperformed more conventional machine learning techniques to natural language processing (Chui et al., 2020). This is primarily because the attention mechanism is combined with a deep learning model based on neural networks. Transformer-based models can generate language knowledge in their neutral networks by retraining their neural networks with large amounts of textual input and using approaches such as learning modeling and sentence prediction. These pre-trained models may then be fine-tuned to fit to a particular application utilizing fresh data. This enables the benefits of deep learning to be applied to far smaller natural language processing (NLP) data sets, which contributes significantly to the models' superior performance (Halim et al. (2020). It is relatively new and has gotten minimal attention from educational institutions, notably in the examination of conceptual knowledge. Many studies have resulted in a range of different techniques to automate comprehension evaluation.

In one research, the latent semantic analysis used the natural language processing approach to compare textual replies to idealized peer responses to provide accurate human grading and forecast post-test performance (Lu Et al. (2021). The discipline of applied NLP, namely in the assessment of conceptual comprehension, is still in its early stages, with few educational institutions adopting it. Another study assessed Learners' understanding of Questions using a combination of natural language processing and node connection representations (Menon et al., 2021). This made it possible to compare commonalities at the knowledge level rather than the textual semantic level and suggests that the advanced technique is capable of determining the pace at which information is reproduced. As a result, a glimpse of learners' conceptual understanding is provided.

Another research established a system for automatically assessing Learners' conceptual comprehension using customized concept inventory questions (Closser et al. (2021). The questions were adapted from the concept inventory of signals and systems, and there was a text answer field where learners could justify their Questions selection (Lin, 2020). NLP approaches were used to determine if a student used an excellent idea and gave sufficient reasoning in their answer. The approach would measure a student's conceptual knowledge level by combining Questions questions with an algorithm that looked for terms suggesting ambiguity. This approach reached a level of accuracy of roughly 85 percent, which is insufficient for classroom application.

Research questions

(1) With this context in mind, our study was directed by the following research questions:

(2) Which natural language processing approaches are most effective for eliciting proof of conceptual comprehension from the text?

(3) The level of performance can be obtained by routinely assessing conceptual knowledge from Learners' textual replies using natural language processing?

(4) What effect does data volume have on the presentation of a conceptual understanding evaluation model that is automated, What does this mean in terms of the future?

2. METHOD

Automated evaluation of conceptual comprehension

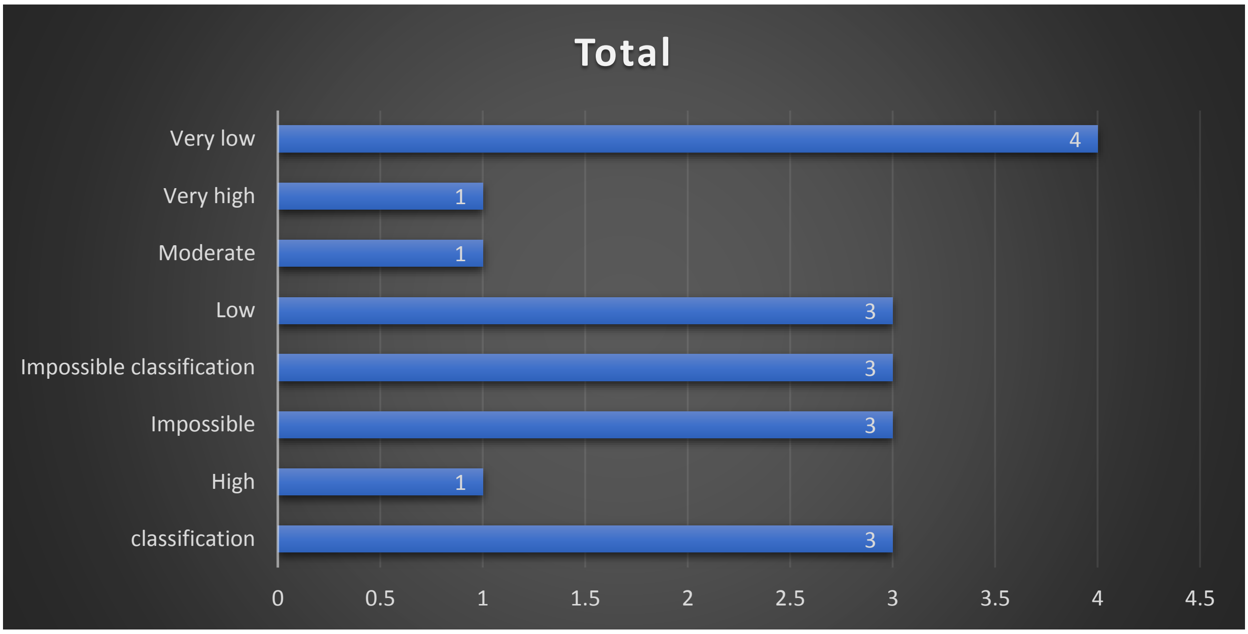
It chose to elaborate on Cunningham-methodologies Nelson's after examining numerous current approaches to automated comprehension evaluation (2019). This is because their study's methodology revealed a great deal about pupils' conceptual knowledge. They might also derive additional information from responses, shedding insight on the nature of pupils' conceptual knowledge, using the increased language comprehension provided by transformer-based NLP models.

Six of the altered idea inventory questions from the previous research are included in the technique created in this investigation (Lai et al., 2020). This research aimed to see if natural language processing methods might be used to evaluate Learners' conceptual knowledge automatically. It chose to build a model that would analyze four pieces of data from a student's response, dubbed pointers, and use them to evaluate the student's conceptual understanding level. Utilizing many leads enables the evaluation of a more complex level of conceptual understanding in literature.

The four points are shown in Table 1 and their binary classifications. The indicator of Response sheds light on the nature of a student's conceptual knowledge.: it indicates how deeply developed their notions are.

Table 1. The four conceptual understanding points and their binary categorization classes are described in detail

Pointer	Description	Classification classes
Questions	If the learner has successfully answered all of the questions in the Questions section	Correct/incorrect
Concepts Free-text validity	Whether or not the student's written argument contains the correct idea or concepts	Yes/No
Free-text validity	Whether the Learners logic is correct and valid in their written justification	Correct/incorrect
Response	The student's level of Response in their written justification	High/low



On a 4.5-point scale ranging from high to poor, an overall model can assess a learner's theoretical comprehension and the presence of misunderstandings utilizing pointer versions capable of automatically detecting each pointer in a response. The process for generating the overall classifications from the pointer classifications is shown in Figure 2.

Figure 2. How categories by pointer affect the total amount of misunderstanding and conceptual comprehension classifications

This dual-output categorization methodology enables instructors to get important information into Learners' conceptual comprehension type and degree. This is advantageous for Instructor since they are often the most difficult areas to handle. Similarly, Instructor can quickly determine whether learners have minimal conceptual comprehension, which is another case in which further work is necessary.

Two logically implausible categories may occur:

- Justification for an adapted notion the inventory question requires the learner to define an idea that is acceptable. When the model determines that the student used appropriate reasoning in their argument without citing the excellent idea or images in their answer, this is judged to be an impossible case.
- If the model determines that the student offered the moral argument but chose the incorrect Questions answer, the condition is declared difficult since a learner who has presented the proper rationale will choose the correct Questions answer.

Comparing the student's decision to the correct answer is a simple way to evaluate the Questions pointer. Additionally, The Concept pointer can be validated by examining the words in a learner's response to a list of idea keywords provided for each individual model list enquiry. NLP modeling will be required to determine the validity of the free-text and response points. Only free-text fact and response leads were examined in this study.

Selection of natural language processing models

It was crucial to identify the most appropriate transformer-based NLP models for this study's application because there are various transformer-based NLP models suitable for a variety of jobs. Differentiating amongst accurate models was helpful in identifying those with the most potential for high performance.

The www.gluebenchmark.com benchmark for general language understanding is a set of data sets that may be used to train, evaluate, and compare natural language processing (NLP) models. The public leader board summarizes the rated models' performance against a human baseline. Due to the diversity of the GLUE data sets.

Pre-processing and collecting of data

Learners' responses were collected for training and testing models during a second-year undergraduate signal analysis course. Between 2015 and 2021, Learners' responses to six modified (Zhao et al. 2020) concept inventory questions were gathered. To ensure that ethical concerns were handled, the questions were presented online as non-compulsory homework.

On the basis of their performance on the GLUE benchmark, its designated models that are best matched and most likely to succeed in determining free-text validity and Response points. To find any differences, manual classifications were used. Figure 3 depicts the numeral of occurrences of each class in the data sets used to answer the question, emphasizing imbalances and data set size.

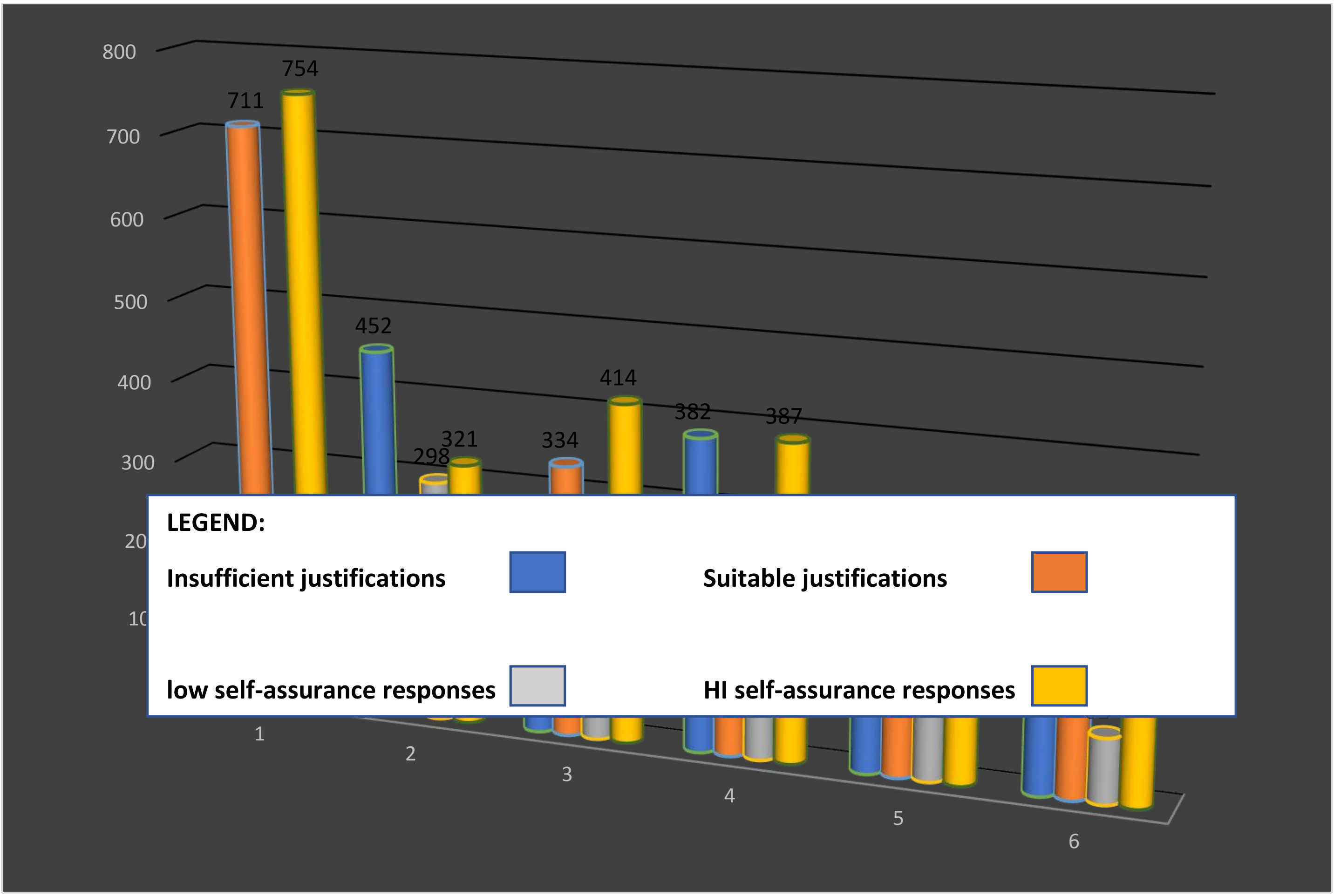


Figure 3. Using both the free-text validity and Response data sets, it was possible to compute the overall number of lesson occurrences per question

To prepare for training, standard pre-processing procedures were used. There was very minimal pre-processing to keep the semantics of Learners' responses and a student's conceptual meaning. This also conforms to criteria for transformer-based models, which do not need extensive pre-processing (Shafiq et al. (2020). All text was converted to lowercase, and punctuation was removed. The Python module spellchecker was then used to spell check and auto correction on misspelled words (Peng et al.,2020). To prevent these terms from being autocorrected wrongly, concept-specific, non-standard-dictionary words must be put into the spellcheck lexicon; examples include Laplace, Fourier, convolution, and Nyquist. Due to the online format of the questions, several duplicate responses were given for each; Learners reattempting the problems in order to find the correct Questions answer most likely caused these.

The data sets were pre-processed to exclude replies consisting of a single word. It was concluded that using a single statement to support a Questions option was improbable. As a result, a one-word response is insufficient for effectively assessing confidence or reasoning. Many single-word replies were invalid, such as a random string of letters. These had been almost certainly the consequence of pupils checking their Questions responses without regard for justification. These tokenizers were chosen and constructed particularly to function with the transformers in question. Numerous Learners supplied reasoning with a question mark after their answer. This demonstrated a lack of confidence in the student's comprehension. As a result, it was agreed that these replies would be automatically categorized as unconfident and excluded from the model training data sets for the Response pointer modeling.

Optimum parameters for model training

The performance of machine learning models fluctuates significantly when training parameters are changed. As a result, a set of model parameters that are appropriate for the given circumstance There are two types of data sets that require NLP models to be defined: Response and free-text validity pointer data sets. Model performance is influenced by the number of batches and epochs utilized in the training process. A stopping delta of 0.05 and a stopping patience of 6 were used as early stopping metrics, with evaluation loss being used as the early stopping metric.

Performance assessment and ensemble modeling

Individual models for each question were trained on separate data sets for the free-text validity pointer model. The accuracy of the receiver operating characteristic curve of the chosen natural language processing types with the ideal training conditions were recorded using an 80:20 split of training and testing for each question. The performance of several natural language processing models may be compared using these findings. The top-performing models were then combined to form an ensemble to develop a model with increased performance. The benefits of each model might then be blended to improve overall accuracy. Because the ensemble model would give a general categorization based on a majority vote (Hew ET al. (2020), the three best-performing models were picked to be combined. Figure 2 depicts the ensemble model.

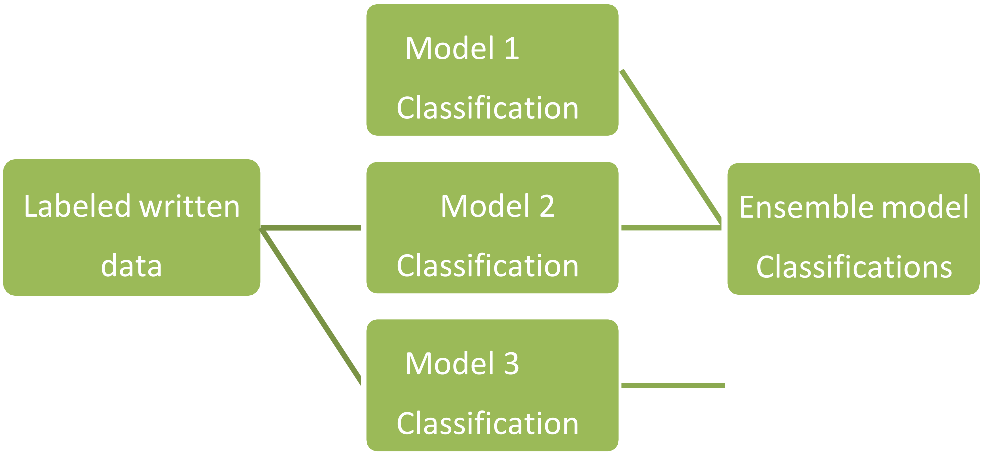


Figure 2. The composition of an ensemble NLP model

Because confidence in the answer is not question-specific, it was determined to build a model that was not question-specific. The ensemble model would provide a general categorization based on a majority vote, The three best-performing models (Hew et al.) were determined to be combined (2020). These findings allow for a comparison of the performance of various NLP models. To increase performance, an ensemble model would be built.

3. RESULTS

Analyzing the GLUE leaderboard

Because the ensemble model would provide a general categorization based on a majority vote, the three most successful models were determined to be combined (Hew et al., 2020). All but the linguistic acceptability corpus required text meaning categorization, according to an examination of the GLUE benchmark data sets. The presentation of the ranking NLP models was investigated in all data sets except the acceptability as a language data set to determine which models would most likely perform well in determining Validity of the free-text. The GLUE benchmark leader board lists four NLP models for sequence categorization:

- ELECTRO
- Roberto
- ALBERTO
- XXLNET

On GLUE data sets that require text classification, the ELECTRO, Roberto, XXLNET, and ALBERTO models outperform the human baseline on average accuracy. Because the data sets used in this study were so small, the models ELECTRO-small, Roberto-base, XXLNET-base, and ALBERTO-V3 were used to simulate free-text validity and Response.

Validity of free-text models

In order to establish the impact of the batch size and epochs training parameters on the free-text validity data set for each question, For the chosen models, a range of values was determined and assessed. It validated the following values:

- The quantity of epochs was increased in increments of one from one to ten, but the number of batches remained constant at eight.
- The number of batches was increased in stages of 50 from 1 to 200, with the epochs set to 1, 5, and 10.

In order to establish the impact of the batch size and epochs training parameters on the free-text validity data set for each question, for the selected models, a set of values was determined and compared. Because of the size of the training data sets used throughout, the number of batches parameter had no effect. Adjusting the number of batches option may become visible when working with more extensive data sets.

The appropriate number of epochs for each chosen NLP model was determined using an early stopping technique. Figure 4 summarizes the effects of premature quitting.

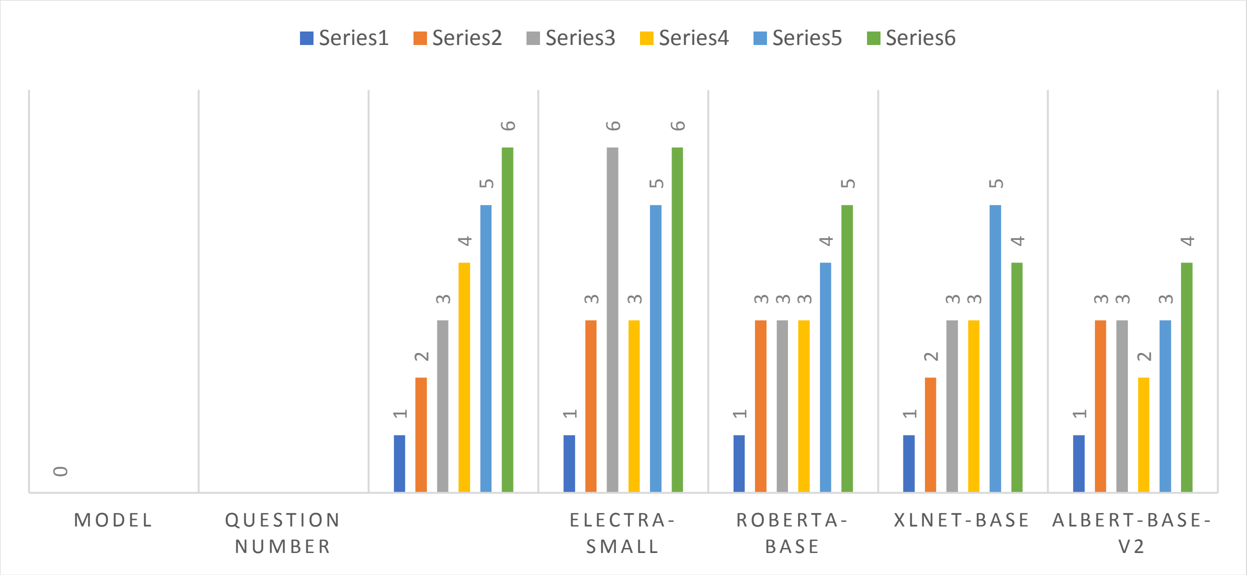


FIGURE 4. The early halting method identified the best number of epochs for each model and questioned the validity of the free-text data set

To see how the batch size and epochs training parameters affected the free-text validity data set for each question, for the chosen models, a range of values was determined and investigated

.

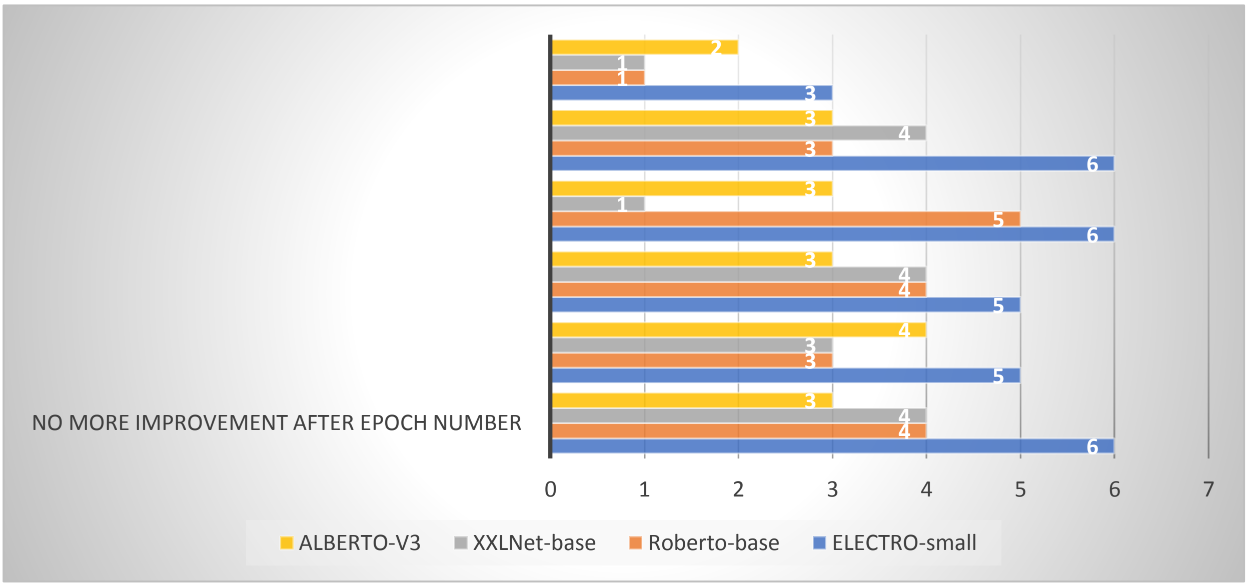


FIGURE 5. The early stopping approach discovered the optimal number of epochs for each model and questioned the set of free-text validity

The ideal number of epochs is discovered to be slightly different when the data from Tables 4 and 5 are compared. Based on the findings in Tables 4 and 5, A a random selection of values for each model was nominated to determine a suitable number of epochs, no matter how large or small the amount of data collected.

Since, ELECTRO-small model appeared regularly in the early halting findings, six epochs were chosen as the optimal value. Figure 6 summarizes the performance of the ELECTRO-small model after six training epochs.

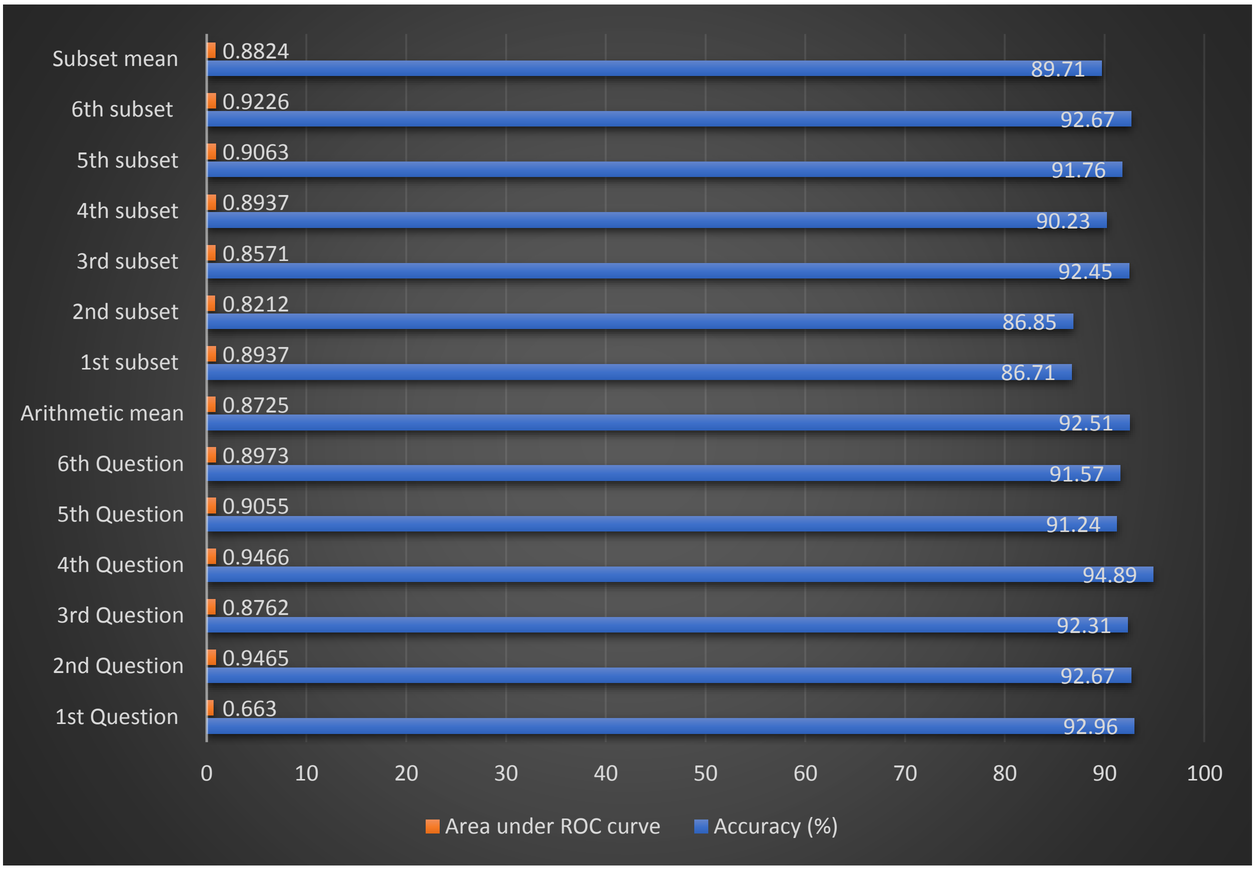
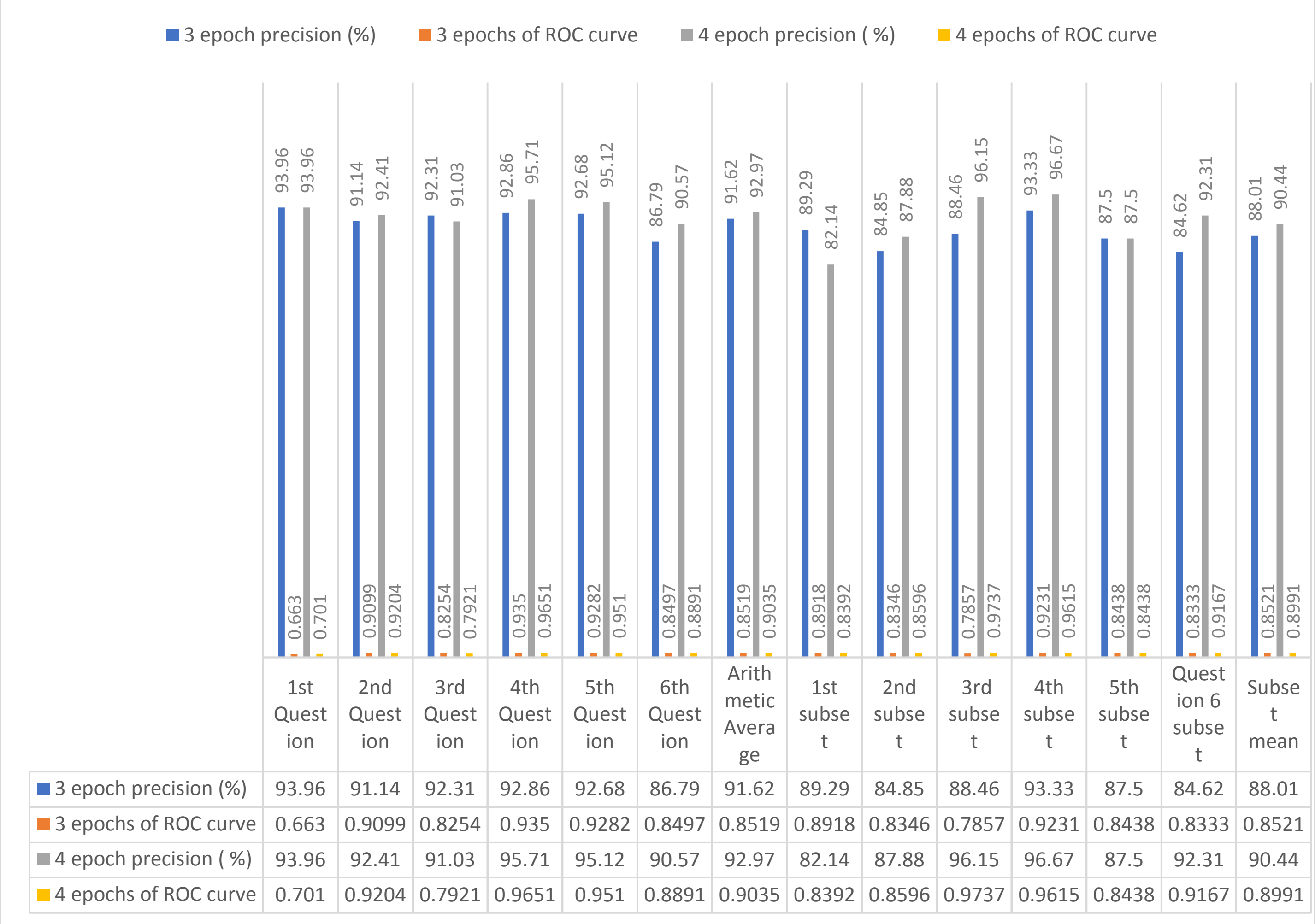


FIGURE 6. The ELECTRO-small model was trained on the free-text validity data set and a subset for each question, and it performed well across six epochs

Three and four epochs were tested in the Roberto-base model based on early halting results. Figure 7 summarizes the performance of the Roberto-based model after three and four epochs of training.



	1st Question	2nd Question	3rd Question	4th Question	5th Question	6th Question	Arithmetic Average	1st subset	2nd subset	3rd subset	4th subset	5th subset	Question 6 subset	Subset mean
3 epoch precision (%)	93.96	91.14	92.31	92.86	92.68	86.79	91.62	89.29	84.85	88.46	93.33	87.5	84.62	88.01
3 epochs of ROC curve	0.663	0.9099	0.8254	0.935	0.9282	0.8497	0.8519	0.8918	0.8346	0.7857	0.9231	0.8438	0.8333	0.8521
4 epoch precision (%)	93.96	92.41	91.03	95.71	95.12	90.57	92.97	82.14	87.88	96.15	96.67	87.5	92.31	90.44
4 epochs of ROC curve	0.701	0.9204	0.7921	0.9651	0.951	0.8891	0.9035	0.8392	0.8596	0.9737	0.9615	0.8438	0.9167	0.8991

FIGURE 7. The performance of the Roberto-based model, trained on the free-text validity data set and each question's subset, over three and four epochs

In the XXLNET-based model, 3 and 4 epochs were selected for analysis due to the early halting findings. Figure 8 summarizes the performance of XXLNET-based models trained during three and four epochs.

Legend

3 epoch precision (%) 🟦

4 epoch precision (%) ⬛

3 epochs of ROC curve 🟧

4 epochs of ROC curve 🟨

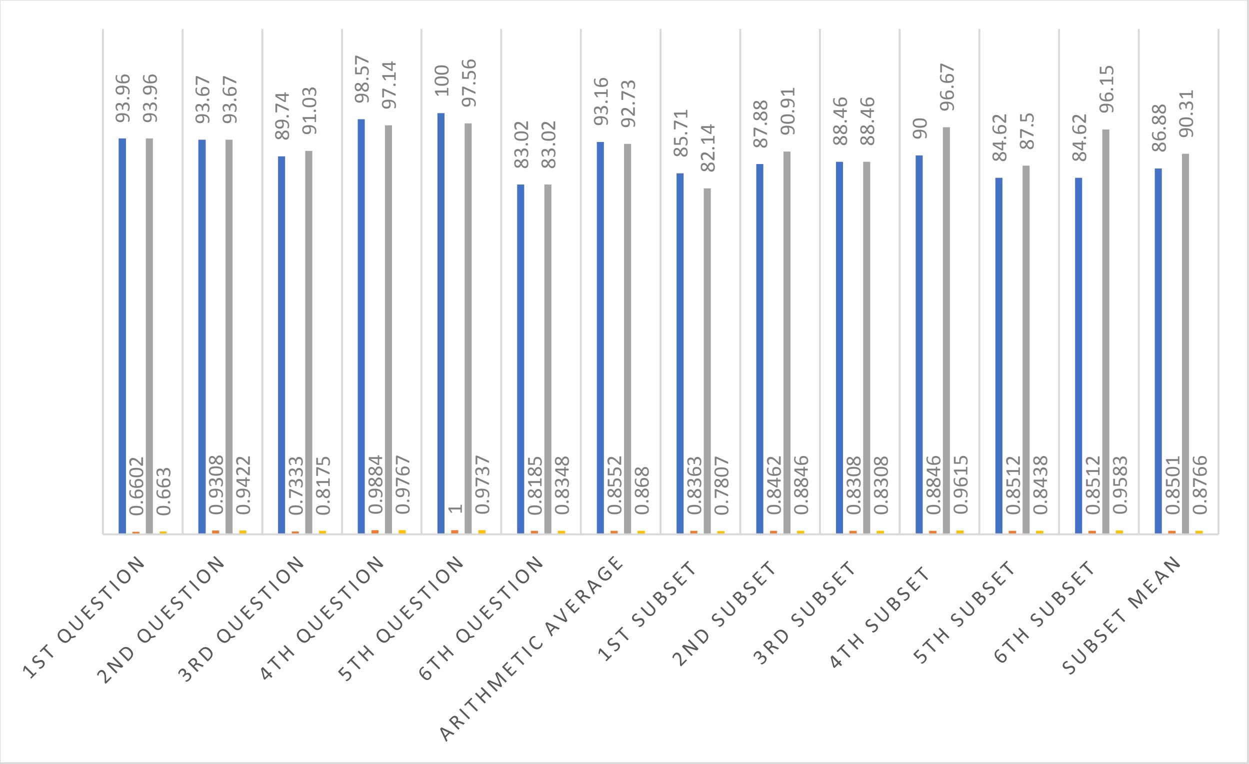


FIGURE 8. Performance of the XXLNET-based model trained on the free-text validity data set and subset for each question with three and four epochs

Three and four epochs were chosen for testing in the ALBERTO-V3 model based on the findings from early halting. Figure 9 summarizes the performance of ALBERTO-V3 models trained across three and four epochs.

Legend

3 epoch precision (%) 🟦 4 epoch precision (%) ⬜

🟧 🟨

3 epochs of ROC curve 4 epochs of ROC curve

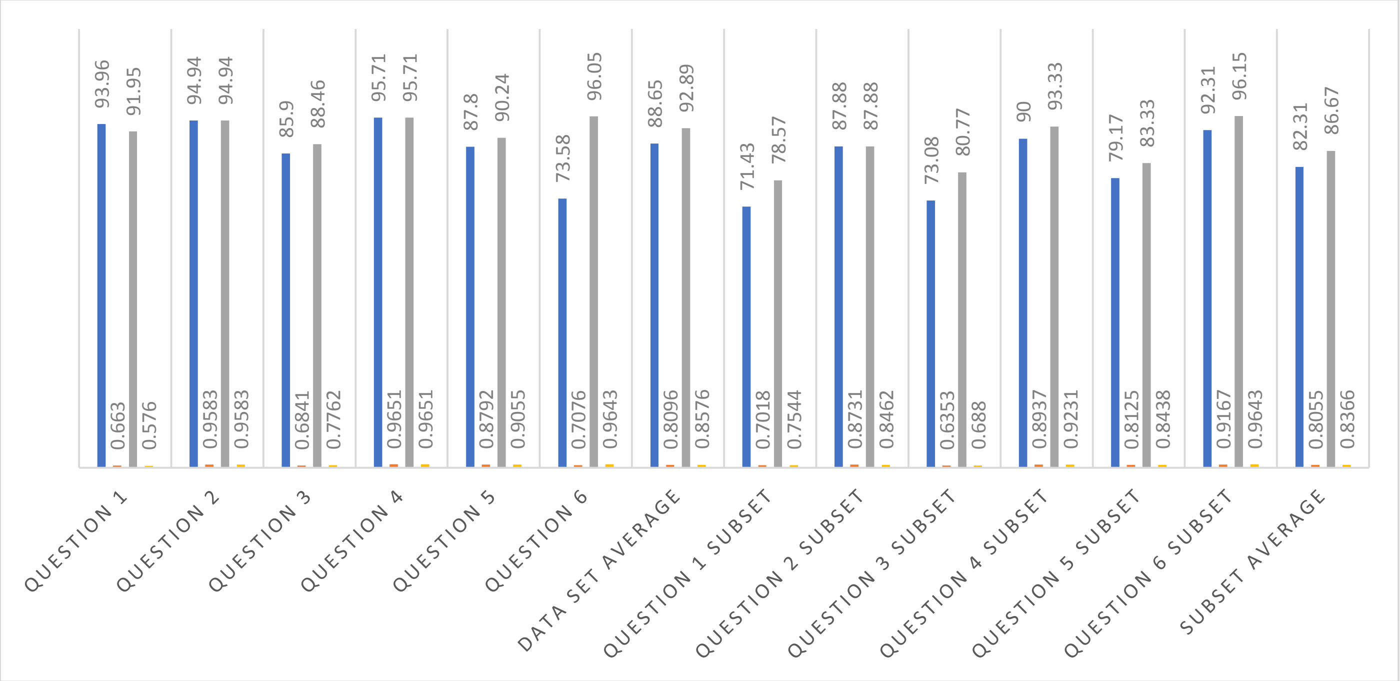


FIGURE 9. The three- and four-epoch ALBERTO-V3 model performance was trained on the free-text validity data set and subset for each question.

Tables 6, 7, 8, and 9 summarize the key findings from the ELECTRO-small model trained with 6 epochs, the Roberto-based model proficient over four epochs and the best results were obtained with a XXLNET-based model trained over four epochs. As a result, a model that combines these three was created. Training the free-text validity ensembles required an average of 35 minutes on a system with an eight-core 16-thread (CPU) and 256(GB)(RAM), reaching from 49 minutes for the most important data set to 30 minutes for the smallest. Figure 10 illustrates the ensemble model's performance on the data set for each topic.

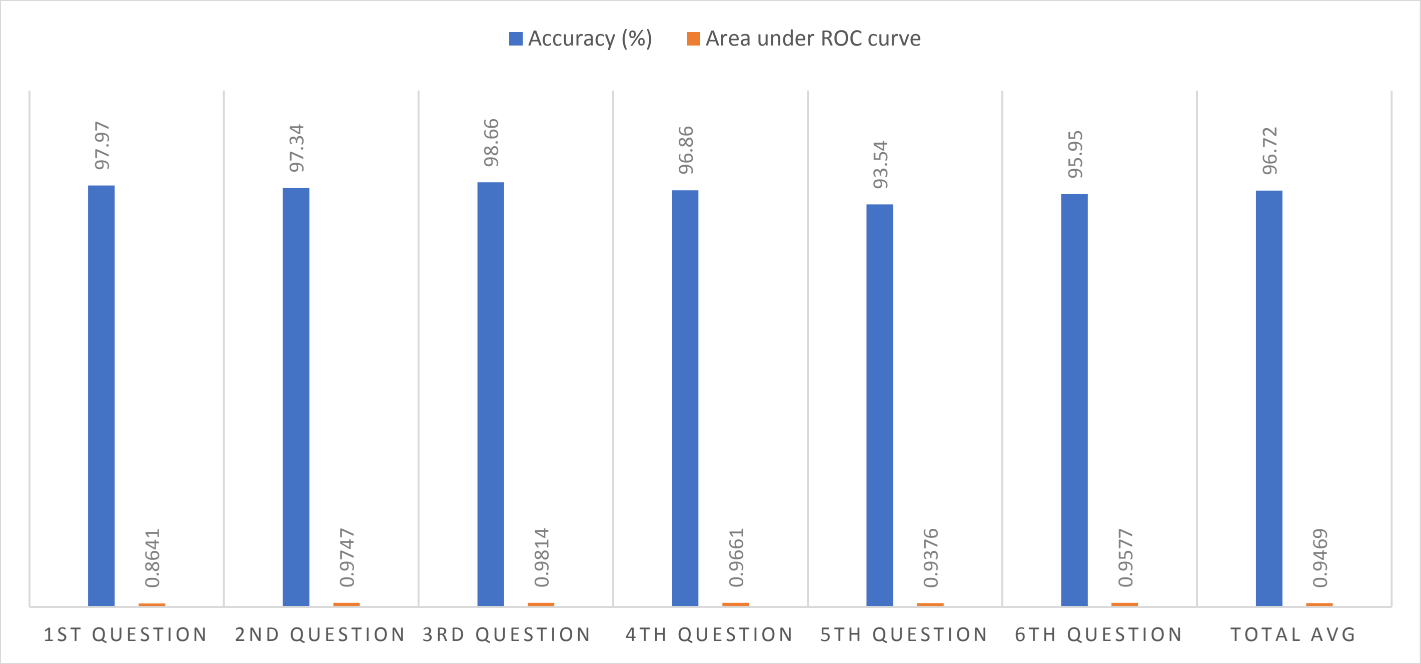


FIGURE 10. Performance of ensemble models with free-text validity on question data sets via fivefold cross-validation

In terms of performance, the ensemble outperformed the individual models. The model could consistently distinguish between effective and faulty reasoning, with a 98.67% average accuracy and a 0.9869 area beneath the curve of receiver operating characteristics. The ensemble model yields really encouraging outcomes, since research shows that Instructor rate their kids' knowledge with less than this degree of accuracy (Elhai et al., 2020).

Performance was tested on two subsets of each question: One contains 100 accurate and incorrect justification responses. The other comprises 40 reasoning responses, both correct and incorrect. to see how well the ensemble applies and adapts to smaller data sets. Tables 11 and 12 indicate how well the ensemble model performed on these two data sets.

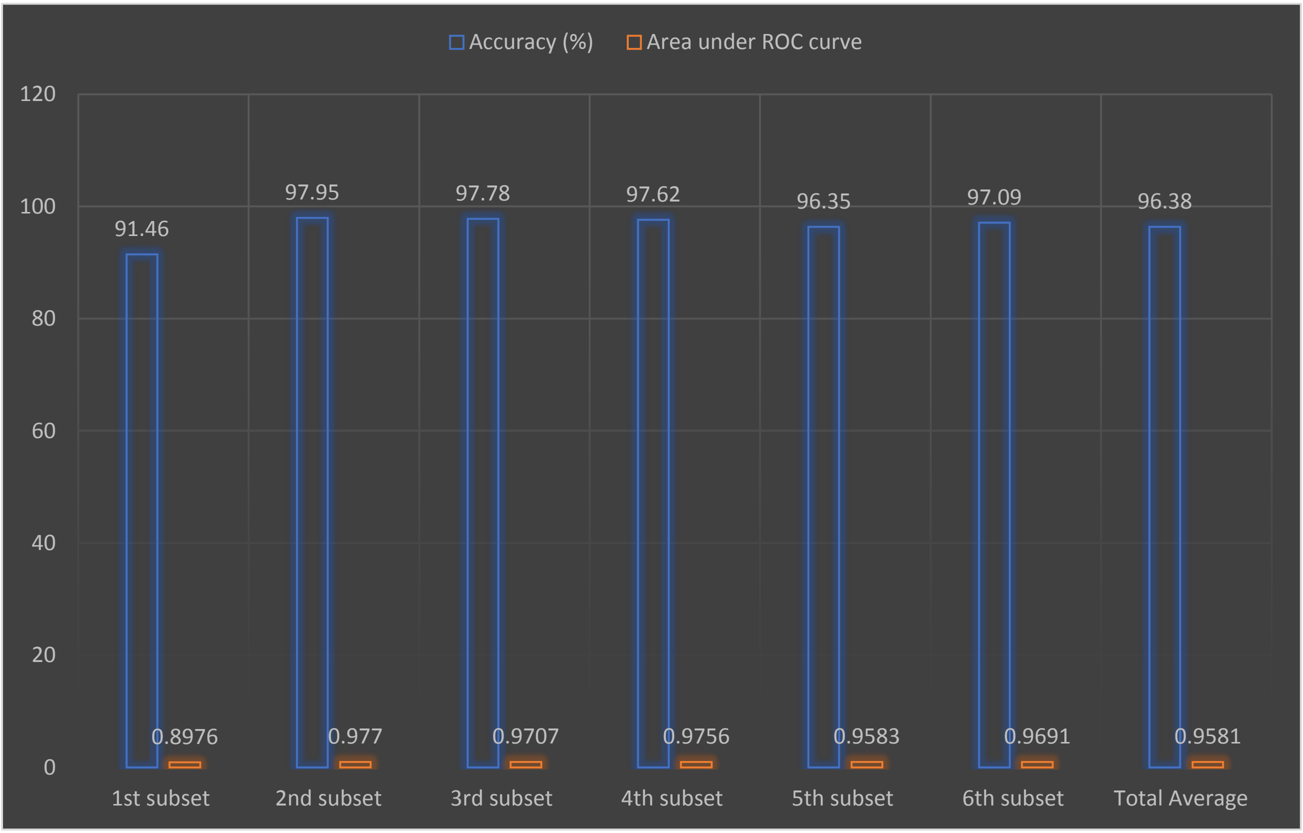


FIGURE 11. Performance of ensemble models on subsets of 100 data sets including valid and wrong justifications

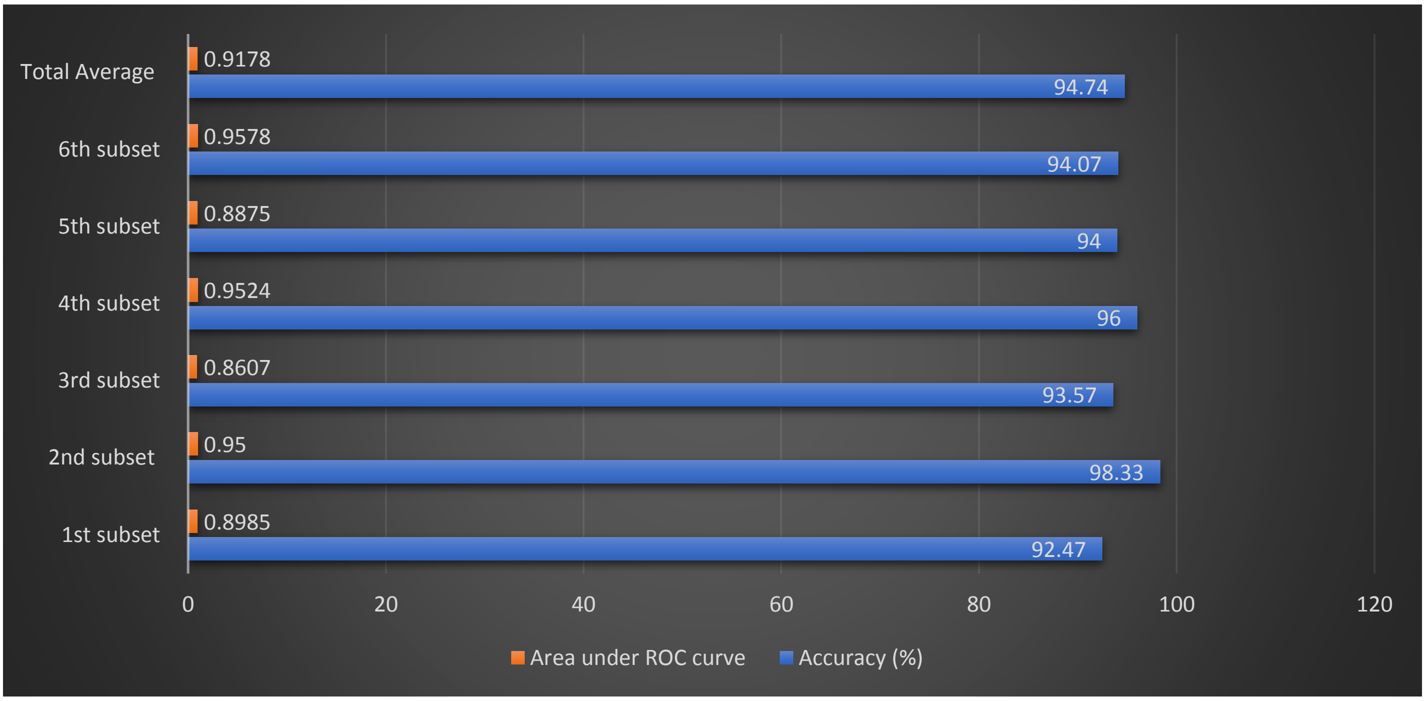


FIGURE 12. Performance of ensemble models on subsets of 40 data sets, including valid and wrong justifications

The ensemble model performed very well, achieving excellent accuracies and areas under the receiver operating characteristic curves across all subset data sets. With only a 0.34 percent loss in average accuracy. Ensembles trained on the complete data set as well as subsets of 100 correct and incorrect reasoning responses yielded results that were almost identical. Additionally, there was a slight decrease in accuracy of 1.98 percent when a substantially smaller group of 40 right and wrong reasoning replies was used. This demonstrates that instructors desiring to incorporate this automated conceptual understanding evaluation technique into their classrooms may do so with little data and get human-like results.

According to Figure 10 On the remaining three questions, the ensemble trained on the whole question data set did the poorest. The first three data sets received a higher percentage of responses than the last three. As a result, models trained on large data sets outperform models trained on smaller data sets. When models are trained on equal student replies, as demonstrated in Tables 11 and 12, performance across questions is equivalent.

Conviction-in-response modeling

Confidence is another crucial indicator that sheds light on pupils' conceptual comprehension quality. Because the batch size parameter had no effect on model performance during free-text validity modeling, it was set at 8. The confidence models' ideal number of epochs was determined using the early halting approach described in the free-text validity modeling. To create the confidence model, the data set for question 1 was picked at random. It was projected that without rebalancing techniques, Due to the considerable imbalance in favor of high self-assurance replies, a serious bias would emerge in the Question 1 data set. Figure 13 summarizes the models' early stopping performance on the training data set.

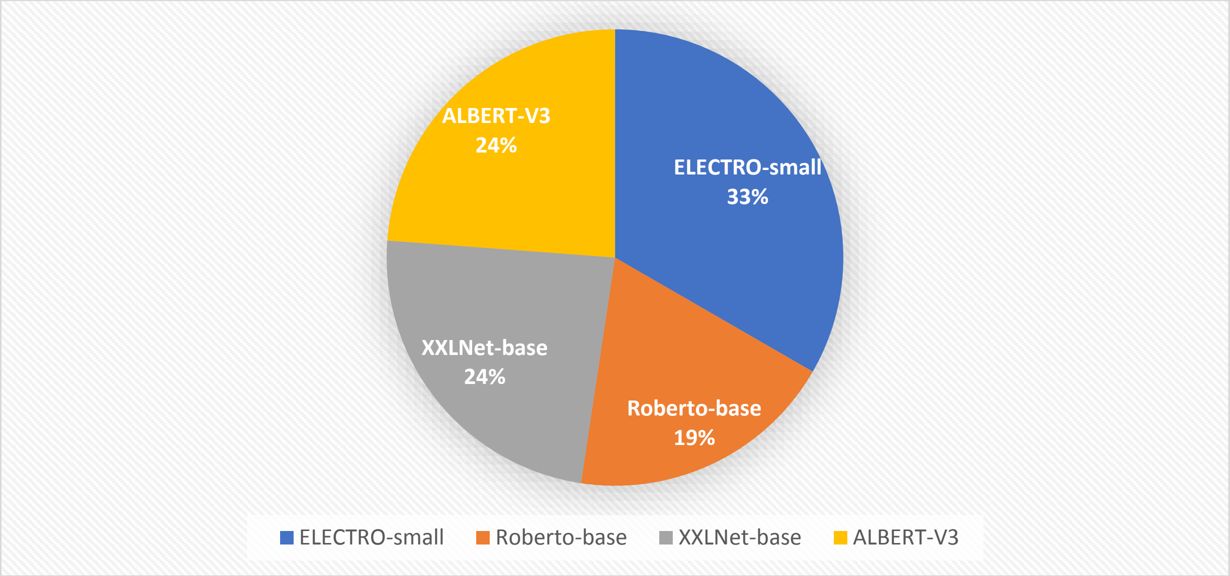


FIGURE 13. On the Question 1 training data set, the early stopping technique found the best number of epochs for each model

The belief in response values for each question was combined to create an extensive testing data set for evaluating the confidence in response models' performance. On the Question 1 training data set, the early stopping technique found the best number of epochs for each model.

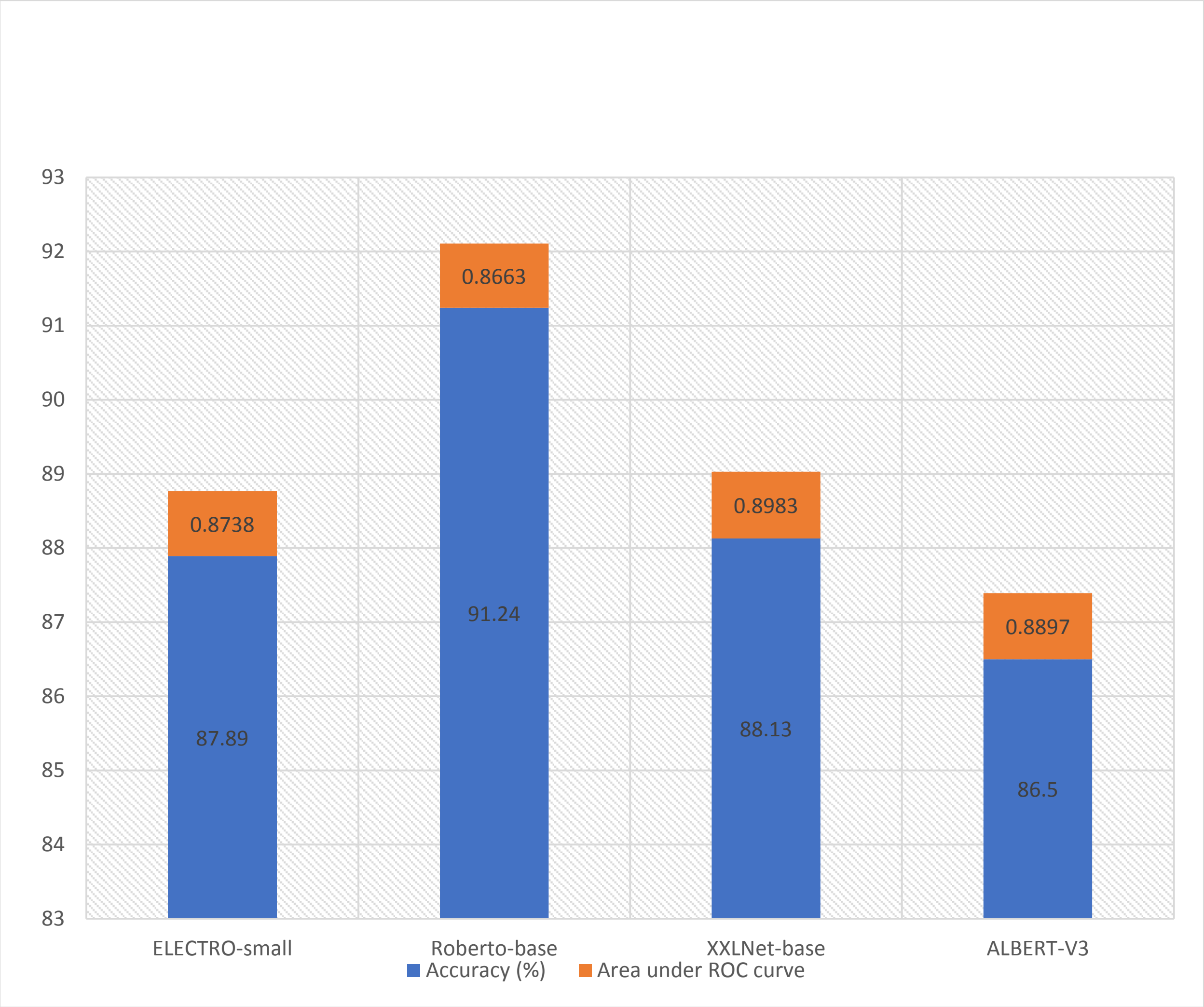


FIGURE 14. On the testing data set, the results of the Question 1 training data set's optimal epoch models

According to Table 14, the results of our performance evaluations, the Roberto-based model qualified over four epochs, the XXLNET-based model qualified over five epochs, and the ALBERTO-V3 model qualified over five epochs all produced the most significant outcomes. As a consequence, an ensemble of these three models was constructed. Despite the fact that the ELECTRO-small model was somewhat more accurate than the ALBERTO-V3 model, It was decided to employ the ALBERTO-V3 model because to its second-highest AUC (area under the receiver characteristic curve). On a PC with an eight-core 16-thread CPU and 256 GB RAM, training this Response ensemble took about 13 minutes. On a PC with an eight-core 16-thread CPU and 256 GB RAM, training this Reaction collective took about 15 minutes.

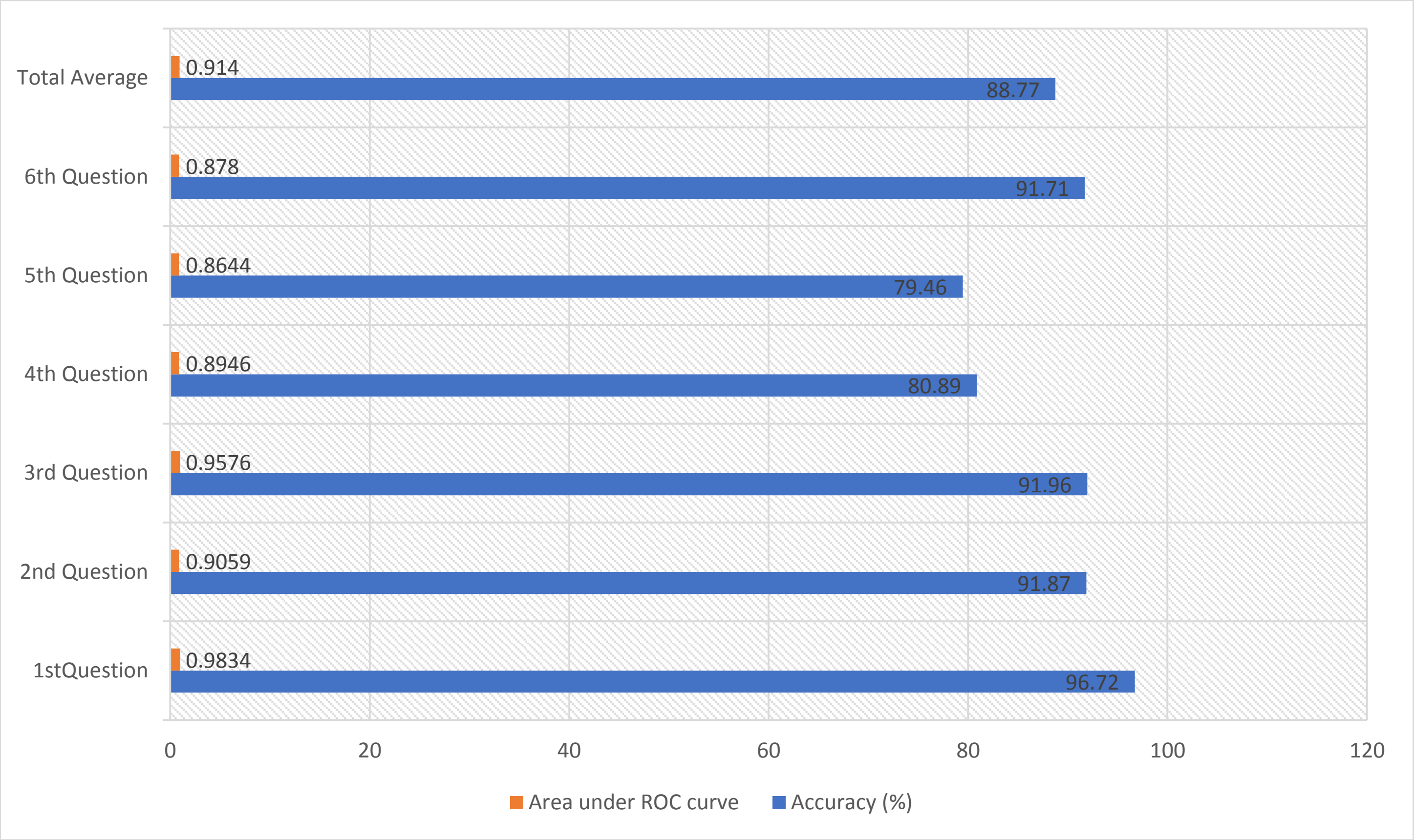


FIGURE 15. Training this Response ensemble took roughly 13 minutes on a system with an eight-core 16-thread CPU and 256 GB RAM

The ensemble model's average accuracy is 88.77 percent, better than the XXLNET- and ALBERTO-V3 models but lower than the Roberto-base model. Figure 15 indicates that, with the exception of challenges four and five, the ensemble performs admirably. When these two searches' misclassifications were reviewed, it was observed that a substantial majority of them contained content-specific, non-dictionary words. The nature of the chosen NLP models could explain the reduction in performance. which rely heavily on retraining language understanding. Despite being pre-trained on large data sets, When the other questions' replies were examined, it was discovered that these content-specific phrases were rarely utilized. The ensemble is unable to understand the words and hence has problems detecting them because it has been competent on replies that do not contain the content-specific phrases. To test this hypothesis and to enhance performance on another question from the original data set, the model was trained using the answers to Question 5 as training data. For each topic, the performance of the ensemble trained on this data set is depicted in Figure 16.

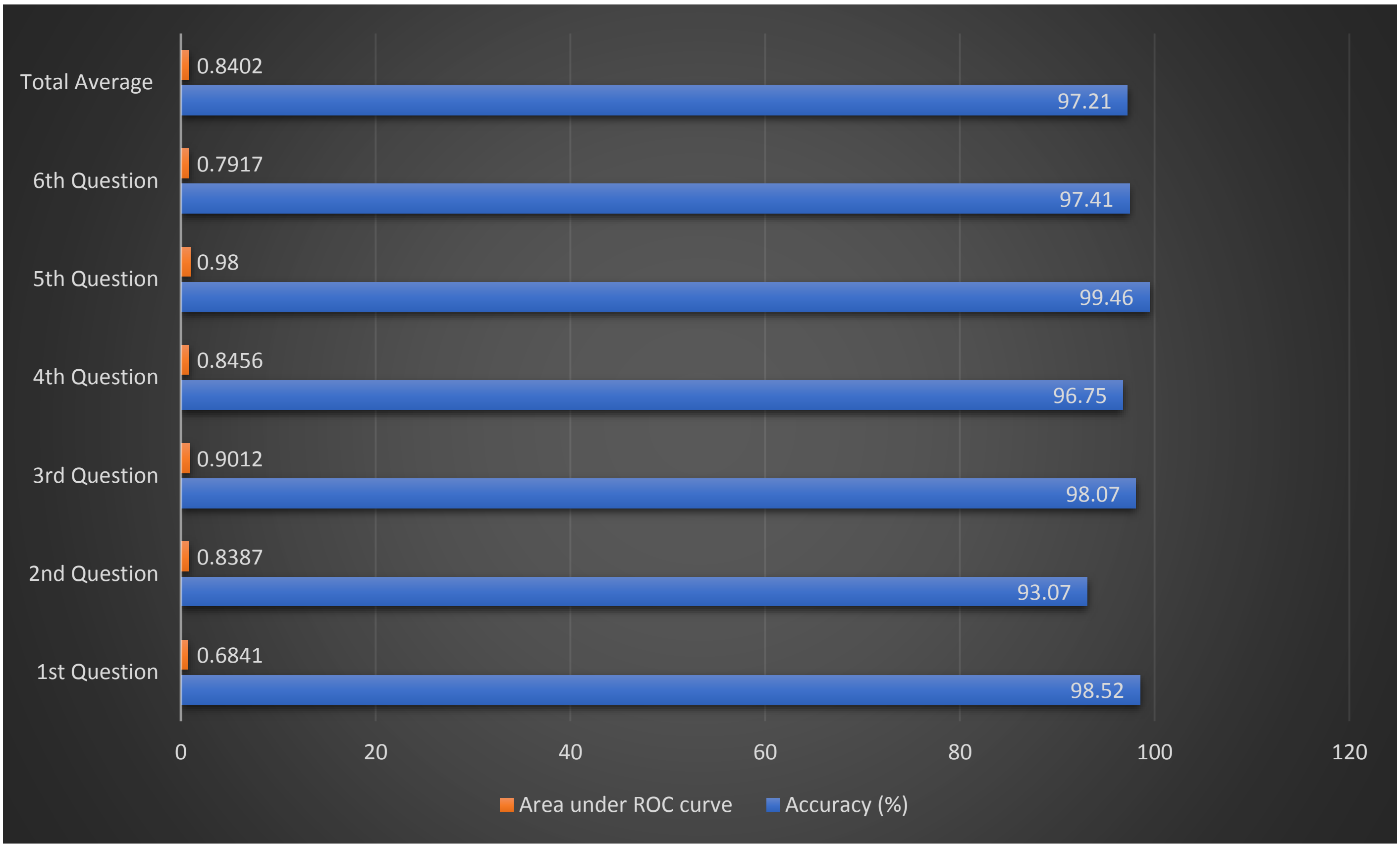


FIGURE 16. After being qualified on the Question 5 training data set, the Response ensemble performed admirably on each data set for each question

This ensemble outperformed the first by a considerable margin. The ensemble could consistently and efficiently distinguish between high and low confidence responses, averages 98.61 percent accuracy and a 0.7421 area under the receiver operating characteristic curve. This performance is superior to that of the ensembles for free-text validity, the ensemble could consistently and efficiently distinguish between high and low confidence responses, having a 97.21 percent average accuracy and a 0.8402 area under the receiver operating characteristic curve. Furthermore, these accuracies are greater than the instructor's ability to gauge Learners' comprehension (Mustafa et al., 2021).

By training the model with the 5th Question training data set, the ensemble's constraints with the Questions 4th and 5th data sets were removed, resulting in considerable presentation improvements. These conclusions propose that pre-trained NLP replicas may struggle with content-specific, non-dictionary terms, but that this can be compensated for during training. Inferring that the study's automated approach produced more consistent and exact results.

Discussion

Modeling NLP can be used to evaluate these two indicators of conceptual understanding in learners' short answer textual responses, according to the validity and response tests for free-text. This is due to the ensemble pointer models' high accuracy, which averages over 97% and has a wide area beneath the receiver operating characteristic curves. When combined with simple algorithms that recognize the other two points, the Learners' Questions selection and the thought sent in explanation, with better than 97 percent accuracy, Automated tools for evaluating conceptual understanding can be created and put to use. The accuracy levels of more than 97% are comparable to instructor evaluations of conceptual knowledge

(Paullada et al., 2021). Thus, Instructor may use the methodologies created in this work as a formative assessment strategy capable of performing to a human level. Additionally, since the method is entirely automated, the approaches may be easily used in flexible and online educational contexts, providing instructors with a new formative assessment alternative with the additional advantage of receiving immediate feedback for themselves and their Learners.

Among the automated evaluation systems examined, the one created in this research gives a unique window into Learners' conceptual knowledge while also delivering more accurate performance. A justification text box, in addition to the conventional Questions concept inventory questions, provides insight into the nature of Learners' conceptual grasp. Furthermore, the method used in this study graded Learners' conceptual knowledge with an accuracy of roughly 85%. (Worsham & Kalita, 2020); our approach significantly outperforms this performance across all points. Our technique also delivers far more information than existing automated understanding evaluation approaches by focusing on conceptual understanding via pointers rather than on resemblance to an example answer (Zhang et al., 2021). Despite the small size of the training data sets, the Response ensembles worked well. These findings suggest that instructors interested in adapting this automated conceptual understanding evaluation technique in their classrooms may do so with little data and obtain good, performance comparable to that of a human. When compared to training a whole transformer model, using pre-trained models reduces training (fine-tuning) time. It also allows for the utilization of prior linguistic expertise.

Additionally, computers with limited memory may be unable to train such models; an 8 GB computer could not learn all ensembles. This is because the transformer models are sophisticated and extensive in size, necessitating a significant amount of computer resources. Fortunately, all ensemble models may be trained online using a variety of free platforms. This implies that instructors may teach their models independent of their available computing resources.

Instructor should be aware that because the study only looked at a limited data set of 90 responses, as a result, to achieve optimal performance, the training data set should include responses that contain content-specific terms. Importantly, teachers should be cautious when expecting content-specific terms in responses, as the Response models performed poorly when the training data lacked such words.

In applications requiring more speed, instructors may choose to consider using more extensive data sets, which would result in improved performance. It is worth noting that data sets with a high degree of class imbalance are predicted to exhibit bias and perform poorly. As a result, it recommends that teachers use a balanced data set for model training when there is a significant imbalance. Furthermore, as this study only optimized the number of batches and epochs, improved performance could be obtained by further tweaking training conditions. As a consequence, where there is a considerable imbalance, it recommends that teachers use a balanced data set for model training. They may be dealt with during pre-processing by the development of an algorithm.

Additionally, other restrictions should be considered in future submissions. The data sets were all derived from the same individual. While the topics used in the questions vary, they are all related to signal analysis. However, this has not been tested. The models would do well in other subject areas. Additionally, additional study is needed to determine how well the models function when dealing with more complicated multi-sentence replies.

4. CONCLUSION

Instructor should be aware that because the study only looked at a limited data set of 90 responses. As a result, to get the greatest performance, replies consisting of content-specific terms should be included in the data for training. Importantly, teachers should be cautious when expecting content-specific terms in responses, as the Response models performed poorly when the training data lacked such words.

In applications requiring more speed, instructors may choose to consider using more extensive data sets, which would result in improved performance. It is worth noting that data sets with a high degree of class imbalance are predicted to exhibit bias and perform poorly. As a result, it proposes that instructors select a balanced data set for model training when there is a significant imbalance. Additionally, enhanced performance may be obtained by further optimizing training settings since this work optimized just the number of batches and epochs. As the study only looked at a small data set with 80 responses. Instructor should keep in mind that both models may sometimes misclassify answers that include many wrong words or phrases. They may be dealt with during pre-processing by the development of an algorithm.

Additionally, other restrictions should be considered in future submissions. The data sets were all derived from the same individual. While the topics used in the questions vary, they are all related to signal analysis. However, this has not been tested. The models would do well in other subject areas. Additionally, additional study is needed to determine how well the models function when dealing with more complicated multi-sentence replies.

5. REFERENCES

1. Broadbent, J., Sharman, S., Panadero, E., & Fuller-Tyszkiewicz, M. (2021). How does self-regulated learning influence formative assessment and summative grade? Comparing online and blended learners. *The Internet and Higher Education*, *50*, 100805. https://doi.org/10.1016/j.iheduc.2021.100805
2. Chen, I. H., Gamble, J. H., Lee, Z. H., & Fu, Q. L. (2020). Formative assessment with interactive whiteboards: A one-year longitudinal study of primary Learners' mathematical performance. *Computers & Education*, *150*, 103833. https://doi.org/10.1016/j.compedu.2020.103833
3. Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university Learners in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*, *107*, 105584. https://doi.org/10.1016/j.chb.2018.06.032
4. Closser, A. H., Erickson, J. A., Smith, H., Varatharaj, A., & Botelho, A. F. (2021). Blending learning analytics and embodied design to model Learners' comprehension of measurement using their actions, speech, and gestures. *International Journal of Child-Computer Interaction*, 100391. https://doi.org/10.1016/j.ijcci.2021.100391
5. El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/10.1016/j.eswa.2020.113679
6. Elhai, J. D., Yang, H., Rozgonjuk, D., & Montag, C. (2020). Using machine learning to model problematic smartphone use severity: The significant role of fear of missing out. *Addictive behaviors*, *103*, 106261. https://doi.org/10.1016/j.addbeh.2019.106261

7. Fernando, W. (2020). Moodle quizzes and their usability for formative assessment of academic writing. *Assessing Writing*, *46*, 100485. https://doi.org/10.1016/j.asw.2020.100485
8. Goularte, F. B., Nassar, S. M., Fileto, R., & Saggion, H. (2019). A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert Systems with Applications*, *115*, 264-275. https://doi.org/10.1016/j.eswa.2018.07.047
9. Granberg, C., Palm, T., & Palmberg, B. (2021). A case study of a formative assessment practice and the effects on Learners' self-regulated learning. *Studies in Educational Evaluation*, *68*, 100955. https://doi.org/10.1016/j.stueduc.2020.100955
10. Halim, Z., Waqar, M., & Tahir, M. (2020). A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-Based Systems*, *208*, 106443. https://doi.org/10.1016/j.knosys.2020.106443
11. Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, *145*, 103724. https://doi.org/10.1016/j.compedu.2019.103724
12. Lin, X. (2020). College student employment data platform based on FPGA and machine learning. *Microprocessors and Microsystems*, 103471. https://doi.org/10.1016/j.micpro.2020.103471
13. Lai, M., Lee, J., Chiu, S., Charm, J., So, W. Y., Yuen, F. P., ... & Zee, B. (2020). A machine learning approach for retinal images analysis as an objective screening method for children with autism spectrum disorder. *EClinicalMedicine*, *28*, 100588. https://doi.org/10.1016/j.eclinm.2020.100588
14. Liu, M., Kitto, K., & Shum, S. B. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior*, *120*, 106733. https://doi.org/10.1016/j.chb.2021.106733
15. Lu, O. H., Huang, A. Y., & Yang, S. J. (2021). Impact of teachers' grading policy on the identification of at-risk Learners in learning analytics. *Computers & Education*, *163*, 104109. https://doi.org/10.1016/j.compedu.2020.104109
16. Ma, J., & Xu, H. (2020). College Learners' network entrepreneurship model based on FPGA and machine learning. *Microprocessors and Microsystems*, 103504. https://doi.org/10.1016/j.micpro.2020.103504
17. Menon, H. K. D., & Janardhan, V. (2021). Machine learning approaches in education. *Materials Today: Proceedings*, *43*, 3470-3480. https://doi.org/10.1016/j.matpr.2020.09.566
18. Martínez-Jiménez, R., & Ruiz-Jiménez, M. C. (2020). Improving Learners' satisfaction and learning performance using flipped classroom. *The International Journal of Management Education*, *18*(3), 100422. https://doi.org/10.1016/j.ijme.2020.100422
19. Otto, H. E., & Mandorli, F. (2021). Advancing formative assessment in

MCAD education: The visual analytics of parametric feature-based solid models. *Advanced Engineering Informatics*, *48*, 101308. https://doi.org/10.1016/j.aei.2021.101308

20. Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2021). Classification and prediction of student performance data using various machine learning algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.382
21. Pallathadka, H., Mustafa, M., Sanchez, D. T., Sajja, G. S., Gour, S., & Naved, M. (2021). Impact of machine learning on management, healthcare and agriculture. *Materials Today: Proceedings*. https://doi.org/10.1016/j.addbeh.2019.106261
22. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336
23. Peng, M., Du, Q., & Zhang, Z. (2020). Children's music teaching visualization system based on FPGA and machine learning. *Microprocessors and Microsystems*, 103510. https://doi.org/10.1016/j.micpro.2020.103510
24. Rahman, S. R., Islam, M. A., Akash, P. P., Parvin, M., Moon, N. N., & Nur, F. N. (2021). Effects of co-curricular activities on student's academic performance by machine learning. *Current Research in Behavioral Sciences*, *2*, 100057. https://doi.org/10.1016/j.crbeha.2021.100057
25. Ray, M. E., DuBrava, L., & Jacks, M. (2020). Leveraging a required e-portfolio course to meet multiple needs: Student assessment, curriculum improvement, and accreditation. *Currents in Pharmacy Teaching and Learning*, *12*(12), 1437-1446. https://doi.org/10.1016/j.cptl.2020.07.004
26. Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research*, *103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602
27. Schellekens, L. H., Bok, H. G., de Jong, L. H., van der Schaaf, M. F., Kremer, W. D., & van der Vleuten, C. P. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation*, *71*, 101094. https://doi.org/10.1016/j.stueduc.2021.101094
28. Shafiq, M., Tian, Z., Bashir, A. K., Jolfaei, A., & Yu, X. (2020). Data mining and machine learning methods for sustainable smart cities traffic classification: a survey. *Sustainable Cities and Society*, *60*, 102177. https://doi.org/10.1016/j.scs.2020.102177
29. Tapingkae, P., Panjaburee, P., Hwang, G. J., & Srisawasdi, N. (2020). Effects of a formative assessment-based contextual gaming approach on Learners’ digital citizenship behaviours, learning motivations, and perceptions. *Computers & Education*, *159*, 103998.

https://doi.org/10.1016/j.compedu.2020.103998

30. Vartiainen, H., Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., & Valtonen, T. (2021). Machine learning for middle schoolers: Learning through data-driven design. *International Journal of Child-Computer Interaction*, *29*, 100281. https://doi.org/10.1016/j.ijcci.2021.100281
31. Veugen, M. J., Gulikers, J. T. M., & den Brok, P. (2021). We agree on what we see: Teacher and student perceptions of formative assessment practice. *Studies in Educational Evaluation*, *70*, 101027. https://doi.org/10.1016/j.stueduc.2021.101027
32. Wu, J. Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and Machine-Learning defined cognitive engagement. *Computers & Education*, *163*, 104066. https://doi.org/10.1016/j.ijcci.2021.100281
33. Worsham, J., & Kalita, J. (2020). Multi-task learning for natural language processing in the 2020s: where are we going?. *Pattern Recognition Letters*, *136*, 120-126. https://doi.org/10.1016/j.patrec.2020.05.031
34. Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., ... & Quintana, R. (2020). eRevis (ing): Learners’ revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449. https://doi.org/10.1016/j.asw.2020.100449
35. Xie, Q., & Cui, Y. (2021). Preservice teachers’ implementation of formative assessment in English writing class: Mentoring matters. *Studies in Educational Evaluation*, *70*, 101019. https://doi.org/10.1016/j.stueduc.2021.101019
36. Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers & Education*, *145*, 103729. https://doi.org/10.1016/j.compedu.2019.103729
37. Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668
38. Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903
39. Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., ... & Sun, M. (2021). CPM: A large-scale generative Chinese pre-trained language model. *AI Open*, *2*, 93-99. https://doi.org/10.1016/j.aiopen.2021.07.001
40. Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society*, *20*, 22-35. https://doi.org/10.1016/j.tbs.2020.02.003