Semi-Supervised Learning Pipeline for Evaluating Question-Answer Pairs Using Generative Pre-trained Transformer 3

Student details
Name: Orçun Özdemir
Student Number: u2086123

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

WORD COUNT: 8780

Thesis committee
Supervisor: dr. Çiçek Güven
Second Reader: dr. Gonzalo Nápoles
External supervisor: Zülküf Genç

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2023

1 Abstract

The utilization of Artificial Intelligence models for the quality assessment of question-answer pairs in Educational Technologies (EdTech) is an area of research that is rapidly expanding, with the potential to improve the quality of educational methodologies, particularly when labelled data is scarce. Consequently, evaluation methodologies of the quality assessment of the question-answer pairs typically require human supervision throughout their pipelines. This thesis focuses on the use of Generative Pre-trained Transformer 3 (GPT-3) models for quality assessment of question-answer pairs without human intervention by employing a Semi-Supervised Learning feedback loop. This thesis consists of two parts: first, generating weakly labelled data with two different equations and creating independent variables by using the relationship between the context of the question-answer pairs and the question-answer pairs themselves. Second, creating two different Semi-Supervised Learning Pipelines (SSLPs) for self-training to evaluate the quality of the question-answer pairs with a limited amount of data. The results demonstrate that the GPT-3 achieves superior results without requiring a large amount of training data and outperforms baseline methods, such as BERT, Random Forest Classifier and Logistic Regression.

2 Ethical Statement

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code.

3 Introduction

Educational Technology (EdTech) is a field that has seen a great deal of application of data science solutions to improve the user experience. It focuses on providing online courses and assessing users with quizzes, making it essential to have high-quality examination resources. There are numerous courses that require an assessment, all of which need question-answer pairs. However, the quality of the question-answer pairs is a major concern, as it is not only about coming up with questions but also ensuring that they assess the right skills and are sound. Traditionally, this requires a lot of non-automatized human evaluation, which is expensive (Nashaat et al., 2018) and, given the amount of data in EdTech, can consume a large portion of the budget of many companies dealing with big data. To save time, money, and human effort, Generative Models are often used, though they still require some manual labeling in their workflow. Additionally, due to the lack of labeled data, Generative Models are used for zero/few/one shot classification to label the data; however, this method is prone

to overfitting, so new models are trained on top of the existing Generative Models. These new models are more accurate since they have been exposed to the problematic data and are less computationally expensive (Smith et al., 2022). Nevertheless, these strategies for improved predictions still need to be evaluated by non-automatized human evaluation.

The Generative Pre-trained Transformer 3 (GPT-3) has been demonstrated to be highly effective in the domain of text data classification (Brown et al., 2020), making it an ideal choice for the text classification objectives of quality assessment of question-answer pairs, which is the primary focus of this thesis. The confidence level of GPT-3 model, as indicated by the conditional probabilities of the outputs when the data is being weakly labelled (See Figure 3), can be utilized as a threshold for generating binary labels which indicate the quality of the question-answer pairs. This is advantageous as it allows for the non-automatized evaluation of the results to be replaced with the confidence level of the token probabilities of the GPT-3 model. This thesis utilizes two distinct GPT-3 models, Curie and Davinci, respectively. Davinci yields superior results in comparison to Curie; however, Curie is faster and more cost-effective (Shihadeh et al., 2022, Hernandez et al., 2021, Zhong et al., 2022). Additionally, the automated labeling is essential for the Semi-Supervised Learning Pipelines (SSLPs), as the weak labels are generated based on the confidence levels prior to the self-training approach.

The evaluation of question-answer pairs has been explored in literature through a variety of strategies, from heuristic rule-based evaluation to taxonomy analysis. However, these strategies have been reliant on human intervention, either in the creation of rules for quality measurement or in the assessment of the classification outputs (Krathwohl, 2002, Smith et al., 2022, Boecking et al., 2021). This thesis seeks to automate the labeling process without human intervention by utilising two self-training approaches using the Weak Supervised Learning. (See Figure 5 and 6)

The Weak Supervised Learning for text classification is a type of Machine Learning technique that utilizes a limited amount of labeled data to train a model. This technique is a form of Semi-Supervised Learning, which is a combination of Supervised and Unsupervised Learning. The Weak Supervised Learning utilizes a small set of labeled data to train a model, and then uses the model to classify the remaining unlabeled data. This approach is beneficial when labeled data is scarce or costly to acquire. Additionally, it can be used to enhance the accuracy of a model by utilizing a small set of labeled data to fine-tune the GPT-3 (Nigam et al., 2000).

The primary objective of this thesis is to develop two Semi-Supervised Learning frameworks for evaluating the quality of question-answer pairs in the EdTech domain. The aim is to predict the quality of the pairs by retrieving the conditional probabilities of the predictions of GPT-3 as a confidence threshold (i.e. good or bad) based on on exploring the association between the context of the question-answer pairs and themselves. To this end, a self-training approach is employed, where GPT-3, Random Forest Classifier and Logistic Regression are utilized, with Bidirectional Encoder Representations from Transformers (BERT)

as a baseline model.

In order to address this primary research query, two subsidiary questions are investigated: What is the most noteworthy independent variable when assessing the features generated from the relation between the context of the question-answer pairs and themselves, such as the Dispersion Score, the Abstractness Score, the Combined Similarity Score, and the Bloom's Taxonomy, for Semi-Supervised Learning to reinforce weakly labeled data for evaluating questions and answers? To what extent the GPT-3 Curie's self-training approach, which utilizes Random Forest Classifier and Logistic Regression, compare to the BERT model as a baseline in terms of transfer learning?

By answering these questions, it should be possible to create a model that not only assesses the quality of question-answer pairs without human intervention but also helps save time and money by using an Semi-Supervised Learning feedback loop.

In order to achieve the primary objective of this thesis, two distinct SSLPs are employed. For both of these pipelines, weak labels are generated with the GPT-3 Curie, do not require human intervention. Both of the SSLPs use the GPT-3 Curie to weakly label the data and the BERT model is adapted for comparison. The first pipeline extracts features, such as the Combined Similarity Score, the Dispersion Score, the Abstractness Score, and the Bloom's Taxonomy, from the relationship between the context of the question-answer pairs. Subsequently, the thesis utilizes the most noteworthy independent variable for the First SSLP, training both a Random Forest Classifier and Logistic Regression model independently to enhance weakly labelled data. Moreover, the Bloom's Taxonomy is a hierarchical system of educational objectives, ranging from the most basic cognitive skills to the highest order thinking skills (Krathwohl, 2002). The Second SSLP involves fine-tuning the GPT-3 Curie with weakly labeled data, which comprises of the weak labels, the context of the question-answer pairs and the question-answer pairs themselves. Both of the pipelines utilize a self-training approach to predict unlabeled data with Semi-Supervised Learning. Subsequently, the performance of both pipelines is evaluated using out-of-sample question-answer pairs, with F1 accuracy scores calculated for Equations 6 and 7.

In the self-training phase of the First SSLP, Random Forest Classifier and Logistic Regression are employed, yielding accuracy results of 64% and 58% respectively when using Equation 6, and 70% and 62% when using Equation 7. In the Second SSLP, GPT-3 Curie is used, achieving an accuracy of 73% with Equation 6 and 85% with Equation 7. To evaluate the accuracy of these pipelines, a fine-tuned BERT model, a deep learning model is developed that leverages a bidirectional transformer architecture to pre-train contextual representations for natural language understanding tasks such as text classification (Devlin et al., 2018), is employed as a baseline model for classifying the quality of the question-answer pairs. The BERT model's accuracy is 73% with the Equation 6 and 74% with Equation 7. The results demonstrate that GPT-3 outperforms the baseline model, Random Forest Classifier, and Logistic Regression on both of the equations, thereby providing an effective solution for assessing

the quality of question-answer pairs.

The out-of-sample evaluation results indicate that the Random Forest Classifier and Logistic Regression with Equation 6 achieved a success rate of 60% and 53%, respectively, while the Random Forest Classifier and Logistic Regression with Equation 7 generated a rate of 65% and 57%, respectively. Subsequently, the GPT-3 Curie's accuracy with Equation 6 and Equation 7 amounted to 67% and 79%, respectively. The BERT model, on the other hand, with the Equation 6 and 7 yielded 65% and 70%, respectively (See Tables 7, 8, 9, and 10).

4 Literature Review

Previous research has sought to address the challenge of effectively classifying text data in the context of question and answer generation. One of the most advanced tools for this purpose is the Generative Pre-trained Transformer 3 (GPT-3) models. GPT-3, an autoregressive language model, has 175 billion parameters and can be applied to any task without gradient updates or fine-tuning, specified solely through text interaction with the model. GPT-3 has demonstrated impressive performance on many Natural Language Processing (NLP) datasets, including translation and question-answering with minimal data (Brown et al., 2020). GPT-3's capacity to answer factual questions is typically approached by using an information retrieval system in combination with a model that can generate an answer given the question and retrieved text. This setting, which allows a system to search for and condition on text that potentially contains the answer, is known as "open-book" (Brown et al., 2020).

GPT-3 has been evaluated using three datasets: Natural, Web, and Trivia Questions. The accuracy for Trivia Questions was 64.3%, 68.0%, and 71.2% in the few/zero/one shot settings, respectively. For Web Questions, the accuracy was 14.4%, 25.3%, and 41.5% in the few/zero/one shot settings, respectively. Lastly, the accuracy for Natural Questions was 14.6%, 23.0%, and 29.9% in the few/zero/one settings, respectively (Brown et al., 2020).This thesis examines the quality of question-answer pairs with a restricted dataset, necessitating the utilization of few/zero/one shot learning. The GPT-3 is underpinned by the Transformer network architecture, which utilises the Attention Mechanism. The Transformer is a neural network architecture that exclusively relies on the Attention Mechanisms, eliminating recurrence and convolutions entirely. The Attention Mechanisms can be described as a mapping of a query and set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is calculated as a weighted sum of the values, with each value's weight being determined by a compatibility function of the query with the corresponding key. The attention mechanism has been applied successfully to a variety of tasks, including text classification, summarization and reading comprehension (Vaswani et al., 2017).

Therefore, this thesis is motivated by the GPT-3, based on the Transformers architecture, which has been demonstrated to surpass state-of-the-art NLP models in terms of text classification when provided with an accurately for-

mulated prompt (Brown et al., 2020, Vaswani et al., 2017, Zhou et al., 2022). Consequently, the GPT-3 is ideal for both the generation of weak labels and the classification of the Bloom's Taxonomy questions. This thesis employs two distinct GPT-3 models, namely Curie and Davinci. Curie is cost-efficient for fine-tuning and faster processing, while Davinci yields the most optimal results and is more accurate in predicting outputs than Curie when the tasks are too complex, albeit at a higher cost. Additionally, Curie is capable of performing nearly as many tasks as Davinci (Shihadeh et al., 2022, Hernandez et al., 2021, Zhong et al., 2022). Therefore, Curie is utilized for self-training and weakly labeling the unlabeled data, while Davinci is employed for the Bloom's Taxonomy Classification in the thesis.

Bloom's Taxonomy is an important theoretical framework being used in this thesis. It is an hierarchical structure for assessing people's ability to learn a given concept, providing definitions for each of the six major categories in the cognitive domain: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation (Krathwohl, 2002). Due to the structure and characteristics of the question-answer pairs utilized in the thesis, Synthesis is not included (see Figure 2). The categories are arranged from simplest to most complex and from most concrete to most abstract; moreover, it is assumed that the Taxonomy represents a cumulative hierarchy, meaning mastery of each simpler category is necessary before moving on to the next more complex one (Krathwohl, 2002). The hierarchical structure of Bloom's Taxonomy allows for the analysis of the average Combined Similarity, Abstractness, and Dispersion Scores of each taxonomy separately. This system of measuring the quality of questions is particularly relevant to this thesis, which focuses on evaluating EdTech question-answer pairs. Consequently, Bloom's Taxonomy inspired the use of it as an independent variable for the First SSLP. Bloom's Taxonomy is a widely-recognized criterion for assessing the quality of questions, yet it only evaluates the quality of questions and not answers. To address this limitation, the hierarchical structure of Bloom's Taxonomy has been employed as a tool for predicting questions' taxonomies with the GPT-3 Davinci and feeding them into the GPT-3 Curie for self-training, which is beneficial for the evaluation of the question-answer pair.

One of the main challenges for this thesis is to adjust the prompt that the models take as input for one/zero/few shot learning predictions. Few-shot learning is a task in machine learning where a model must learn from a limited number of examples, which is often difficult for traditional machine learning models that require large amounts of data in order to learn. Recent progressions in pre-trained language models have indicated that they can be utilized for few-shot learning tasks by fine-tuning the models (Gao et al., 2021). It is an important leap forward in the domain of machine learning, as it suggests that pre-trained language models can be used to learn from a limited number of samples, which is usually a challenging situation for conventional machine learning models. (Gao et al., 2021) The success of the pre-trained model in the few-shot setting has inspired this thesis to investigate the potential of fine-tuning the GPT-3 Curie in a zero-shot setting within the Second SSLP (See Figure 6). In particular, the efficient utilization of limited data is advantageous to the thesis in terms

of making the most of the available data resource. It is suggested making predictions on an existing pre-trained model instead of fine-tuning a blank model (Gao et al., 2021, and this thesis implements the process of fine-tuning on top of an existing pre-trained model, GPT-3 Curie.

This thesis proposes the use of token probabilities of GPT-3 Curie as a confidence metric for labelling data, particularly in the context of Semi-Supervised Learning with weak labelling. Weak labelling is a method of training a model using only partially labelled data, which is often employed when there is an insufficient amount of labelled data to train a model using traditional methods. This approach has been demonstrated to be effective in various settings, such as text classification (Meng et al., 2018, Chapelle et al., 2009). Semi-Supervised Learning is often utilised when there is a limited dataset, particularly with regard to the number of labelled samples (Meng et al., 2018, Chapelle et al., 2009). Consequently, this thesis is motivated by the potential of Semi-Supervised Learning to efficiently evaluate small datasets due to the scarcity of labelled data.

In order to compare the GPT-3's performance with a state-of-the-art language model, the BERT model is utilized. BERT model BERT is a language model which utilises Transformers with encoder layers and self-attention heads. It differentiates itself from other standard language models by having a bidirectional Transformer architecture. To overcome the unidirectional constraint, they employed a Masked Language Model (MLM) (Devlin et al., 2018). The objective of the MLM is to randomly mask some of the tokens from the input and predict the original vocabulary id of the masked word-based solely on its context. This objective enables the depiction to fuse the left and right context, thus allowing for the pre-training of a deep bidirectional Transformer. In addition to the masked language model, the "next sentence prediction" task is also used to jointly pre-train text-pair representations (Devlin et al., 2018). It is a state-of-the-art tool for gathering the most semantically similar sentence and word-based comparison with the question. Furthermore, it is particularly useful for feeding large texts, such as question-answer pairs, which are the inputs of this thesis. BERT models do not require labelled data to match semantically similar scores due to its bidirectional encoder representations. Consequently, the lack of labelled data available for this thesis aligns with the BERT design in this regard.

Additionally, machine learning models have been successfully managed with weak supervision, as the large amount of data can cause bottlenecks in the training phase (Boecking et al., 2021). A framework, Interactive Weak Supervision, that receives user feedback to propose heuristic solutions, is developed and demonstrated that even with a small amount of intervention from user feedback, it is possible to achieve competitive results without accessing ground truth labels. The weak labels are created under human supervision and fed into an Artificial Neural Network (ANN) model (Boecking et al., 2021). What inspired this thesis is the fact that they fed the supervised labels into the ANN model for reinforcement. Although GPT-3 is a highly advanced generative AI model, it is evident from the weak labels it produces that it needs to be self-trained, as in the Second SSLP (See Figure 6). A heuristic rule-based prompt has been

proceeded and used a set of rules previously created by humans (Boecking et al., 2021). This thesis, on the other hand, does not take a heuristic approach, neither for the independent variable extraction in the First SSLP nor for the GPT-3 Curie's self-training in the Second SSLP.

In addition, A survey-like approach is employed to prompt creation for spam classification (Smith et al., 2022). Subject Matter Experts (SME) formulated questions that were posed to pre-trained language models, with the expectation of receiving yes or no answers. The results were then evaluated heuristically; if the model responded affirmatively to the query of whether the text requested an action, it was flagged as spam (Smith et al., 2022). Subsequently, labeled data was fed into an ANN and used to predict the unlabeled data. SME then reviewed the predicted models in a feedback loop. This approach is similar to that of this thesis in terms of utilizing large language models with prompts and creating a feedback loop to improve predictions. However, this thesis's probabilistic filtering approach and heuristic approach of the previous research differ. Moreover, SME have been necessitated to make the final decision (Smith et al., 2022), whereas this thesis focuses on the capacity of Generative Models for final predictions. This thesis seeks to employ a strategy for quality assessment through prompt with GPT-3, which is inspired by the text classification using prompt addressed by the previous literature. However, a heuristic approach is employed to address the challenge, with the outputs being evaluated by human experts and the results being adjusted according to predetermined criteria (Smith et al., 2022). In contrast, this thesis does not take an external heuristic approach, both for the independent variable extraction in the First SSLP (See Figure 5) and for the self-training of the GPT-3 Curie in the Second SSLP (See Figure 6).

5 Data

The main input for evaluation is quizzes which are consisted of a question, and multiple choice answers. The quizzes are generated with zero-shot learning approach by using GPT-3 Davinci. The prompt consists of the description of the generation task, sample input and output format, and a context that retrieved from video transcripts of Skillsoft courses. Skillsoft is a software company that provides online learning solutions to businesses and organizations (Skillsoft, 2022).

After feeding the prompt with given context, the GPT-3 model generates quizzes for evaluation. Subsequently, there can be multiple true and false answers in one quiz, and the unlabeled data contains 1446 different quizzes from 992 different video transcripts and each quizzes contains 5 to 9 different question-answer pairs. In order to apply the methodologies, the thesis separate each question,good and bad answers and extract their logarithmic probabilities individually.

The GPT-3 models utilise logarithmic probabilities to determine which words to complete a sentence with. Logarithmic probabilities are a method of statis-

Figure 1: Sample of Question-Answer Pairs, X_i

`Context:`
Informative context about vaccines.
`Question:`
Which of the following statements is an example of an informed judgment about the value of vaccines?
`Correct Answer:`
Vaccines are beneficial because they prevent diseases that can cause serious harm or death to individuals and populations.
`Incorrect Answers:`
— Vaccines are harmful because they introduce foreign substances into the body that can cause allergic reactions or autoimmune disorders.
— Vaccines are unnecessary because natural immunity is stronger and more reliable than artificial immunity.
— Vaccines are controversial because they are based on unproven theories and lack sufficient evidence of safety and effectiveness.

tically estimating the likelihood of a specific word being selected (Brown et al., 2020). The intricate description of the logarithmic probabilities is presented in Section 6.2.

Due to the scarcity of labeled data, this thesis proposes a weakly labeled dataset to initiate self-training. To address this challenge, this thesis leverages generated question-answer pairs from the GPT-3 Davinci for initial weak labeling. There are 130 different generated quizzes, each containing 3 to 5 different question-answer pairs, for a total of 634. Subsequently, these question-answer pairs are evaluated with the GPT-3 Curie to retrieve their logarithmic probabilities using the following formula:

$$prob_i, Y_i, tokenY_i = GPT3(X_i) \tag{1}$$

where $prob_i \in R_n$ is the corresponding logarithmic probabilities of n tokens, Y_i is a textual binary output, $tokenY_i$ is the token list of the Y_i and the input of the formula containsX_i. It is the question-answer pairs and their related context (See Figure 1). The context is the transcripts which are used when the question-answer pairs are generated using the GPT-3 Davinci.

6 Methodology

The initial step in this process is the prompt creation for one/zero/few shot learning. Subsequently, independent variables such as Bloom's Taxonomy Label, Combined Similarity Score, Dispersion Score, and Abstractness Score are established. Two SSLPs for Generative Models are then set up to binary classify the quality of question-answer pairs. The First SSLP (See Figure 5) utilizes independent variables for a self-training approach, while the Second SSLP (See

Figure 6) uses a fine-tuned GPT-3 Curie for self-training. This study employs a one/zero/few shot learning approach and a key factor in this strategy is the preparation of an appropriate prompt for the GPT-3 Davinci and Curie. After retrieving the question-answer pairs and their contexts, they are fed to two different SSLPs.

The First SSLP starts with three steps. Firstly, independent variables, namely the Abstractness Score, the Dispersion Score, and the Combined Similarity Score, are created from the relation between the question-answer pairs and their contexts. This step is essential for the First SSLP, as it feeds the Random Forest Classifier and Logistic Regression. Secondly, the Bloom's Taxonomy of the input question is classified by using the GPT-3 Davinci. Thirdly, the question-answer pairs are labeled with the GPT-3 Curie based on its token probabilities of predictions. After these steps are completed, another dataset is created consisting of the newly created features and weakly labeled dependent variables. Subsequently, unlabeled data from Skillsoft is retrieved and a self-training operation is initiated. During the training, the Random Forest Classifer and Logistic Regression predict and evaluate the unlabeled data in a loop (see Figure 5).

The Second SSLP begins with two steps. Firstly, weakly labeled dependent variables are created using the same technique as the First SSLP. Secondly, instead of creating independent variables for self-training, transfer learning of the GPT-3 Curie is initiated. In this step, question-answer pairs and their contexts are fed into a JSON file to send an API request to OpenAI servers. The servers then return the fine-tuned model with an identifier, allowing access to and utilization of the fine-tuned GPT-3 Curie for prediction. Following these steps, as with the First SSLP, unlabeled data from Skillsoft is retrieved and the fine-tuned GPT-3 Curie is re-trained in a loop until the unlabeled data is labeled (see Figure 6). Further subsections explain the creation of independent and dependent variables, and classification algorithms in detail.

6.1 Creating Independent Variables

6.1.1 Bloom's Taxonomy Labels

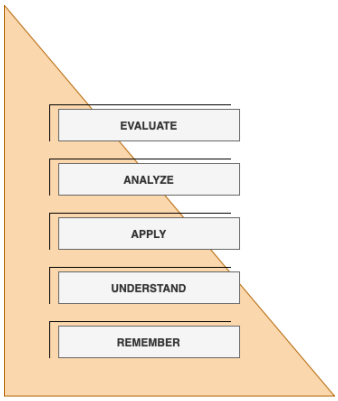


Figure 2: Bloom's Taxonomy

One of the features of this thesis is the classification of questions based on Bloom's Taxonomy (See Figure 2). For the First SSLP (See Figure 5), another approach similar to Formula 1 is employed to classify the Bloom's Taxonomy of the input question. In this case, Y_i in the Formula 2, is a textual output with five labels: Understand, Apply, Analyze, Evaluate, and Remember. Bloom's Taxonomy is a classification system used to describe the different levels of thinking and learning that are involved in the educational process (Anderson and Krathwohl, 2001). The labels of Understand, Apply, Analyze, Evaluate, and Remember are the five levels of the taxonomy. Understanding involves the ability to comprehend the meaning of a concept or idea. Applying involves the ability to use the concept or idea in a new situation. Analyzing involves breaking down a concept or idea into its component parts and examining the relationships between them. Evaluating involves making judgments based on criteria and standards. Finally, Remembering involves the ability to recall information. Together, these five levels of thinking and learning provide a framework for educators to use when designing instruction and assessing student learning (Anderson and Krathwohl, 2001). Furthermore, the prompt used for Formula 1 has been modified to taxonomy classification, rather than assessing the quality of the question-answer pairs.

This study employs a one/zero/few shot learning approach to classify Bloom's Taxonomy. A key factor in this approach is the preparation of an appropriate prompt for the GPT-3 Davinci. This prompt is a string description which contained the definitions of Bloom's Taxonomy and its levels, sample context, questions, and the taxonomy level of the question. For classification, contexts and questions to be classified are fed to the prompt dynamically, with the outcome predicted by the Bloom's Taxonomy level of the input question. GPT-3 Davinci is chosen for this task due to its ability to handle the more complex prompt for classification of the Bloom's Taxonomy compared to GPT-3 Curie's prompt for creating weakly labels. The equation for classifying the Bloom's Taxonomy using the GPT-3 model is as follows:

$$Y_i = TaxonomyClassifierGPT3(X_i) \tag{2}$$

Let Y_i denote the Bloom's Taxonomy of the input Question, which comprises Understand, Apply, Analyze, Evaluate, and Remember. X_i is the prompt that contains the context related to the question-answer pair, as well as the

question-answer pair itself. The $TaxonomyClassifierGPT3$ function is used to feed the prompt to the GPT-3 Davinci, which then predicts the Bloom's Taxonomy of the given question. Finally, GPT-3 Davinci predicts around 49% of questions that are classified as "Remember", 16% of questions classified as "Understand", less than 1% of questions classified as "Apply" and "Analyze", and 23% of questions classified as "Evaluate" to be used in the First SSLP (See Figure 5).

6.1.2 Combined Similarity Score: Semantic Similarity and Word-Based Similarity Scores

The other feature that is employed for weak supervised learning is the combination of semantic similarity and word-based similarity scoring. The inputs for these methods are questions and contexts of the questions. As it is explained in the section 4, this method utilises BERT model for semantic similarity matching between the context and the question. Subsequently, the lack of labelled data available for this thesis aligns with the BERT design in this regard. The combined score is calculated as follows:

$$CS_{ij} = \frac{SimilarityScore(ContextSentence_j, Question_i)}{(1 + FuzzyScore(ContextSentence_j, Question_i))} \tag{3}$$

The Combined Similarity Score is evaluated by comparing each sentence of the context ($ContextSentence_j$) with the input question ($Question_i$) using the BERT model. The most semantically similar pairs are retrieved with the Similarity Score function, and then the same pairs are compared with the Fuzzy Score function and summed with one to determine if the most semantically similar sentences are also word-wise similar. The two scores are then divided to obtain the Combined Similarity Score, CS_{ij}. This score is intended to penalise questions that are almost exactly replicated in the text, as this may lead students to memorise the sequence rather than actually learn it.

6.1.3 Dispersion Score

The thesis focuses on another feature, which is calculating the entropy among the top N semantically similar sentences. After retrieving the sentences, their scores are collected and the entropy are calculated based on the similarity scores. This approach is developed to comprehend the dispersion of the sentences within the input text. The formula for this feature is as follows:

$$DispersionScore_i = -\sum_{j=1}^{N}(p_j * log(p_j)) \tag{4}$$

The Dispersion Score of a given $question_i(DispersionScore_i)$ is calculated by performing matrix multiplication between p_j and the logarithm of p_j, and then taking the negative of the result. This score is based on the scores of the N

most semantically similar sentences of the given $context_i$ of the question-answer pairs and the $question_i$.

6.1.4 Abstractness Score

Another feature for weakly supervised learning is the Abstractness Score. The formula of the Abstractness Score is the following:

$$AS_{ij} = \frac{SimilarityScore(ContextSentence_j, Question_i)}{1 + \frac{1}{5}\sum_{n=2}^{7} CommonNGrams_n(ContextSentence_j, Question_i)} \quad (5)$$

The Abstractness Score(AS_{ij}) is calculated by comparing each sentence of the context ($ContextSentence_j$) with the input question (Question_i) and taking their n-grams. This process is similar to the Combined Similarity Score, where the Semantic Similarity Score is retrieved with the $SimilarityScore$ function. The Semantic Similarity Score is a measure of the degree of similarity between the context and the correct answer(s). The methodology for calculating the Semantic Similarity Score is analogous to the Combined Similarity and the Dispersion Scores. Five different n-grams, ranging from two to seven, of the most semantically similar $ContextSentence_j$ and $Question_i$ are taken with the $CommonNGrams$ function and averaged. Then, the output of the $CommonNGrams$ function is summed with one in the event that there are no common n-grams.

Finally, the division of the Semantic Similarity Score and the summed average score is calculated and set as an independent variable for the model. The purpose of creating such a feature is to determine if the correct answer is present in the context directly. Furthermore, this feature is used to ascertain whether the model created its answers directly from the context or if it generated a certain abstraction to the answer. Generating answers directly from the context is not desirable, as it typically does not assess students' understanding but rather their memorization ability.

6.2 Creating Weakly Labeled Dependent Variables

In order to initiate a training process, weak labels of question-answer pairs must be generated and this is accomplished by using logarithmic probabilities of tokens from Formula 6 or Formula 7.

The GPT-3 Curie calculates the probability of a certain word following a given sequence of words. It then uses the logarithmic probability of the most likely word to complete the sentence. The logarithmic probability is calculated by taking the logarithm of the probability of the word divided by the sum of the probabilities of all other words. In other words, it is the logarithm of the probability of the word given the context of the sentence. This allows the model to determine the most likely word to complete the sentence, even in cases where there are multiple possible words to choose from (Brown et al., 2020).

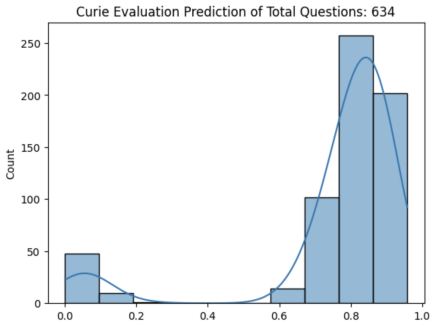


Figure 3: GPT-3 Curie's Token Probabilty Histogram

The generated question-answer pairs from GPT-3 are retrieved as a dictionary, thus each question, true and false answers must be extracted separately. After the tokens and token probabilities of each are retrieved and compared with one another, two different approaches are implemented for retrieving weak labels. Firstly, the probabilities of each question, correct and false answers are extracted separately and summed using the following formula, where P_{tokenY_j} is the conditional logarithmic probability of a token in each sentence and t_j is the token.

$$ConfidenceScore = \sum_{j=1}^{n} exp(P_{tokenY_j}(t_j|t_1, ..., t_{j-1})) \tag{6}$$

Subsequently, since most of the correct and incorrect answers are multiple choice, a Softmax function is employed to normalize them. This ensures that the sum of correct and incorrect answers is equal to one, respectively. Histograms of each question and its corresponding correct and incorrect answers (See Figure 3) are then generated and a threshold is assigned for labeling. If the probability is below the given threshold, it is labeled as weak; otherwise, it is labeled as strong. Furthermore, labels are retrieved from the evaluation data based on the probability of each question-answer pair. In this case, each word from the question-answer pairs is collected and their average probability is calculated. Consequently, the average probabilities of each question, correct answers, and incorrect answers are retrieved separately using the following formula, where P_{tokenY_j} is the conditional logarithmic probability of a token in each sentence and t_j is the token.

$$ConfidenceScore = \frac{1}{n} \sum_{j=1}^{n} exp(P_{tokenY_j}(t_j|t_1, ..., t_{j-1}))) \tag{7}$$

Due to the limited amount of labeled data available, supervised learning cannot be effectively implemented without the risk of over-fitting. To address this issue, this thesis proposes the use of self-training learning. Self-training is

a Semi-Supervised Learning technique that involves training a classifier with a small set of labeled data, then using the classifier to label unlabeled data. The most confident unlabeled points, along with their predicted labels, are added to the training set and the classifier is retrained. This process is then repeated (Zhu, 2005).

6.3 Classification with Weakly Labeled Data

This thesis investigates two distinct self-training approaches. The first approach entails predicting weakly labeled question-answer pairs using independent variables with Random Forest Classifier and Logistic Regression (See Figure 5). The second approach involves predicting weakly labeled question-answer pairs with prompt based classification using a fine-tuned Curie based on its token's conditional probabilities (See Figure 6).

Moreover GPT-3 Curie is fine-tuned on the evaluation dataset and used to predict unlabelled data. Three datasets are used for this operation: a train and test set from the evaluation data, and unlabeled data for the machine and deep learning models to label with self-training. The train and test data contain weak labels generated from Curie's token probabilities.

GPT-3 Curie is first fine-tuned with the train and test data, then the evaluation data is predicted and its F1 scores are retrieved. The unlabelled data is then predicted, resulting in both labels and probabilities of the labels. Labels with probabilities higher than 90% are fed into the train data and trained again until there is no more unlabeled data.

In contrast to the First SSLP, which utilises the strategy of creating weak labels with independent variables such as the Abstractness Score, the Dispersion Score, the Bloom's Taxonomy, and the Combined Similarity Score, the Second SSLP employs Curie's state-of-the-art word embeddings. For this method, instead of using the independent variables for predicting unlabelled data (See Figure 5), the context and the question-answer pairs are fed to GPT-3 as a prompt and are expected to be input for the self-training pipeline (See Figure 6). Subsequently, in the prompt, classification of the question-answer pairs is asked whether they are good or bad. Therefore, instead of using scores as independent variables, the conditional probability of token, which is generated by GPT-3, is utilised. According to the histogram (See Figure 3) of the sample data for evaluation, whenever GPT-3 Curie generates a strong label with 95% token probability of a good question-answer pair, it is assumed that it is indeed a good pair. GPT-3 Curie is proficient at retrieving high probability of high quality questions; however, low probability of high quality questions are also needed.

In order to accomplish this objective, when GPT-3 Curie is not certain in classifying question-answer pairs with a confidence score of less than 20% for a good question-answer pair (See Figure 3), it is assumed to be a negative pair. This approach utilizes fine-tuning, thus the model is not supplied with extra independent variables. Ultimately, GPT-3 Curie's probabilities on predictions

are analysed. This also reduces the disparity between pre-training and fine-tuning, making it more functional in few shot scenarios (Gao et al., 2021).

7 Evaluation

In order to evaluate the primary question of the thesis, question-answer pairs are weakly labeled and the weakly labeled data is split into test and train data, and fed into both of the SSLPs. The F1 scores of the pipelines are then retrieved to assess the validity of GPT-3's confidence threshold based on the association between the context of the question-answer pairs and themselves. To address the significant generated independent variable with the relationship of the context of the question-answer pairs and themselves, feature importances of each independent variable are calculated and evaluated. To investigate the performance of GPT-3 Curie's self-training compared to Random Forest Classifier, Logistic Regression, and the baseline model, a fine-tuning operation is implemented. Thereafter, the GPT-3 Curie in the Second SSLP is fine-tuned using a transfer learning approach.

In this thesis, a train-test split operation of Scikit-learn (Pedregosa et al., 2011) is conducted on the 634 question-answer pairs, with the default parameters used apart from the train and test sizes, which are configured to 70% and 30% respectively. Additionally, the training data comprises of around 80% good and 20% bad labels, and the test data containes approximately 75% good and 25% bad labels. As a result of the imbalanced data, F1-score is utilized as the accuracy metric.

The fine-tuning of GPT-3 Curie is elucidated in the following: two objects must be inputted into the models, the prompt and the labels. The prompt for fine-tuning is distinct from the prompts that are utilized for Bloom's Taxonomy classification and quality classification of the question-answer pairs. In this prompt, there is no directive, only the independent variables, similar to fine-tuning a BERT model. Subsequently, the labels consist of good or bad. Both of the inputs are collected in JSON format and transmitted via API request to OpenAI. Subsequently, OpenAI stores the fine-tuned model in their repository and the models are accessible via their unique identifiers.

In order to assess independent variables, the average scores of Abstractness, Dispersion, and Combined Similarity are obtained based on the taxonomy labels. This approach facilitates the observation of the efficacy of the scores. For instance, it is anticipated that a Remember type of question's Combined Similarity Score is lower than an Understand type of question's Combined Similarity Score. Subsequently, the importance of features is compared based on their impact on the quality of the question and the most significant ones are shared. To achieve this, feature importance based on feature permutation (Pedregosa et al., 2011) is used for each machine learning algorithm. After retrieving each feature's importance (See Figure 4, the most important ones for both Equation 6 and 7 are selected.

After generating the features for the final component of the pipeline, Random

Forest Classifier and Logistic Regression are employed to classify the question-answer pairs for deployment using the generated features. During this stage, the classifiers are evaluated based on F1 accuracy scores. This study focuses on binary classification, which indicates whether the question-answer pairs are good or bad.

Furthermore, the BERT model (Devlin et al., 2018) is used as the base model with transfer learning. To accomplish this, the BERT model is fine-tuned with the same training data as the self-training models, with question-answer pairs as independent variables and good/bad labels as dependent variables. Then, the fine-tuned BERT model is assessed using the same test data as the self-training models. Finally, all of the self-training models and the BERT model are benchmarked based on their F1 accuracy scores.

8 Result

In the methodology section, two different weak labeling strategies are discussed. One of these strategies is the average of logarithmic probabilities, which can be seen in Figure 3. This See Figure demonstrates a clear saturation between low confidence and high confidence predictions. Additionally, GPT-3 Davinci classifies Bloom's Taxonomy labels with given questions and context (Table 1). The table indicates that the overall micro F1 accuracy scores for Remember, Understand, Apply, Analyze, and Evaluate Labels are 64%. It is observed that Understand type of questions are predicted weakly, as the difference between Understand and Remember type of questions is quite small. In many use cases, the difference between the two is that Remember type of question's words are mostly in the context. Therefore, the model struggles to differentiate between the two taxonomies. To examine this further, the Combined Similarity, the Abstractness, and the Dispersion Scores were separated by their Bloom's Taxonomy labels (Table 2). The average Combined Similarity Scores of the quizzes are as follows: Remember type of question is 40.9%, Understand type of question is 66.7%, Analyze type of question is 53.5%, Apply type of question is 64.5%, and Evaluate type of question is 53.4%. This indicates that word-based similar sentences with the question of each quiz penalize the score. Consequently, Remember type of questions clearly obtain low scores with this formula, since the question's words are more present in the text than any other question types.

Table 1: Accuracy Results of Bloom's Taxonomy Prediction by GPT-3 Davinci

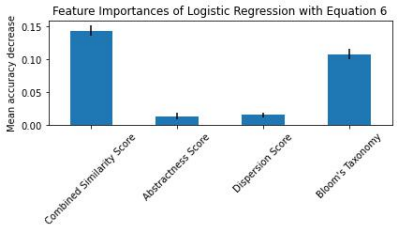
Taxonomy	Precision	Recall	F1-score	Support
Remember	0.82	0.85	0.84	312
Understand	0.60	0.38	0.46	104
Apply	0.50	0.80	0.67	57
Analyze	0.42	0.54	0.49	13
Evaluate	0.72	0.76	0.74	148
Overall	0.61	0.66	0.64	634

Subsequently, the average Dispersion Scores of the quizzes reflect similar outcomes to the combined similarity scores. For the Remember, Understand, Apply, Analyze, and Evaluate type of questions, the average Dispersion Scores are 86.3%, 90.2%, 90.9%, and 89.4%,85.9% respectively (Table 2). The Dispersion score indicates the density of the distribution of the similarity scores between context's sentences and questions, and the percentages match with their Bloom's Taxonomy labels accordingly. It is logical that the Dispersion Score would be higher in Remember type of questions since semantic similarity algorithms are more likely to detect word-based similarity. Additionally, the average abstractness scores of the quizzes based on their taxonomies indicate how correctly answers are mentioned in the context. To evaluate students taking the quiz more accurately, the choices should not be directly in the context which can lead students to memorize the answers. Therefore, the abstractness score measures the abstractness of correct answers. The average Abstractness Scores of the quizzes are 31.7%, 40.2%, 51.5%, 37.8%,and 50.4% for Remember, Understand, Analyze, Evaluate, and Apply type of questions respectively (Table 2). The results indicate that as the taxonomy level increases hierarchically (Krathwohl, 2002), the values of the scores also increase.

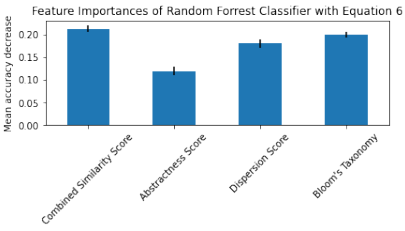
Table 2: Average Abstractness, Combined Similarity, and Dispersion scores of each Bloom's Taxonomy Labels

	Average Abstractness	Average Combined Similarity	Average Dispersion
Remember	0.317	0.409	0.863
Understand	0.402	0.667	0.902
Analyze	0.515	0.535	0.894
Apply	0.504	0.645	0.909
Evaluate	0.378	0.534	0.859

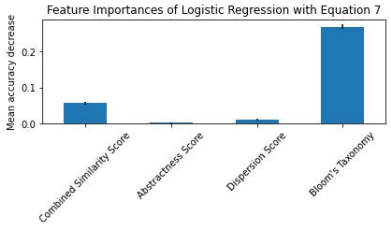
To answer the first research sub-question, after analyzing the feature importance (Pedregosa et al., 2011) of Random Forest Classifier and Logistic Regression, it was determined that the Combined Similarity Score is the most significant feature for the labels generated from Equations 6 and 7. Consequently, the Combined Similarity Score is found to be the most relevant independent variable to context and question-answer pairs due to its semantic and word-based similarity scores for the Equation 6. Subsequently, the Bloom's Taxonomy is determined to be the most pertinent independent variable in regards to context and question-answer pairs because of its informative qualities when applied to the questions at hand.



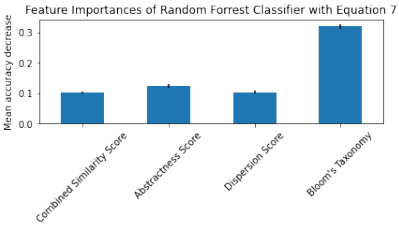
(a) Feature Importance of Logistic Regression with Equation 6



(b) Feature Importance of Random Forest Classifier with Equation 6



(c) Feature Importance of Logistic Regression with Equation 7



(d) Feature Importance of Random Forest Classifier with Equation 7

Figure 4: Feature Importance of Random Forest Classifier and Logistic Regression with Equation 6 and 7

The accuracy scores of the first SSLP (See Figure 5) are retrieved for two different data labeling equations. Subsequently, F1 accuracy scores of predicting labels generated with Equation 6 on test data with Random Forest Classifier, Logistic Regression, and BERT are 64.3%, 58.2%, and 72.8%, respectively (Table 3). Additionally, F1 accuracy scores of predicting labels generated with Equation 7 on test data with Random Forest Classifier, Logistic Regression, and BERT are 70.1%, 61.7%, and 73.7%, respectively (Table 4). The second SSLP's (See Figure 6) accuracy scores are also retrieved for two different data labeling equations. F1 accuracy scores of predicting labels generated with Equation 6 on test data with GPT-3 Curie and BERT are 73.7% and 72.8%, respectively (Table 5). Subsequently, F1 accuracy scores of predicting labels generated with Equation 7 on test data with GPT-3 Curie and BERT are 85.3% and 73.7%, respectively (Table 6).

Results of predicting labels using Equation 6 are more accurate with the Random Forest Classifier than Logistic Regression in the first SSLP (See Figure 5), and in certain simulations Logistic Regression fails to label all of the unlabelled data in the dataset. The reason for this failure is that Logistic Regression, after a few iterations, cannot produce predictions whose probabilities are greater than 90%. As a result, approximately 20% of the unlabeled data remains unlabeled with low probability predictions. On the other hand, Random Forest Classifier managed to predict 90% of the unlabeled data. For the labels which are created with Equation 7, the machine learning algorithms fol-

low a similar pattern. Again, the Random Forest Classifier is able to predict approximately 80% of the unlabelled data and Logistic Regression is able to predict approximately 60% of the unlabeled data. To answer the second research sub-question, in the second pipeline (See Figure 6), GPT-3 Curie is able to predict approximately 80% of the unlabeled data generated with Equation 6. Subsequently, GPT-3 Curie is able to predict approximately 90% of the unlabeled data generated with Equation 7. Additionally, the F1 accuracy score of the state-of-the-art BERT model was 73.7%, which is close to that of the First SSLP (See Figure 5), even though both the BERT model and the Second SSLP (See Figure 6) employed a similar fine-tuning strategy. The reason for this is that the BERT model was only fine-tuned once, whereas GPT-3 Curie in the Second SSLP was kept fine-tuned until there are no unlabeled data left.

When predicting the out-of-sample data, Logistic Regression achieved an accuracy of 53% while Random Forest Classifier achieved an accuracy of 60% when using Equation 6. Moreover, when Equation 7 is applied to create labels, Random Forest Classifier is able to predict the out-of-sample data with an accuracy of 65%, and Logistic Regression achieved an accuracy of 57%. Additionally, GPT-3 Curie is able to predict the out-of-sample data created using Equation 6 with an accuracy of 67%, and 79% when Equation 7 is used. Finally, BERT model is able to predict the out-of-sample data created using Equations 6 and 7 with accuracies of 65% and 70%, respectively (See Tables 7, 8, 9, and 10).

Table 3: Accuracy Results of First SSLP (See Figure 5) with the Equation 6

Algorithm	Precision	Recall	F1-score
RFC	0.663	0.607	0.643
LOGIT	0.562	0.602	0.582
BERT	0.734	0.712	0.728

Table 4: Accuracy Results of First SSLP (See Figure 5) with the Equation 7

Algorithm	Precision	Recall	F1-score
RFC	0.684	0.732	0.701
LOGIT	0.602	0.643	0.617
BERT	0.753	0.722	0.737

Table 5: Accuracy Results of Second SSLP (See Figure 6) with the Equation 6

Algorithm	Precision	Recall	F1-score
GPT-3 Curie	0.705	0.793	0.737
BERT	0.734	0.712	0.728

Table 6: Accuracy Results of Second SSLP (See Figure 6) with the Equation 7

Algorithm	Precision	Recall	F1-score
GPT-3 Curie	0.851	0.822	0.853
BERT	0.753	0.722	0.737

Table 7: Out-of-Sample Accuracy Results of First SSLP (See Figure 5) with the Equation 6

Algorithm	Precision	Recall	F1-score
RFC	0.566	0.582	0.600
LOGIT	0.485	0.511	0.532
BERT	0.639	0.620	0.651

Table 8: Out-of-Sample Accuracy Results of First SSLP (See Figure 5) with the Equation 7

Algorithm	Precision	Recall	F1-score
RFC	0.639	0.623	0.650
LOGIT	0.603	0.591	0.574
BERT	0.725	0.718	0.702

Table 9: Out-of-Sample Accuracy Results of Second SSLP (See Figure 6) with the Equation 6

Algorithm	Precision	Recall	F1-score
GPT-3 Curie	0.646	0.634	0.672
BERT	0.639	0.620	0.651

Table 10: Out-of-Sample Accuracy Results of Second SSLP (See Figure 6) with the Equation 7

Algorithm	Precision	Recall	F1-score
GPT-3 Curie	0.801	0.763	0.793
BERT	0.725	0.718	0.702

9 Discussion

In order to evaluate the effectiveness of question-answer pairs in Edtech, two distinct SSLPs are suggested. The first approach involves extracting and comparing the given question-answer pairs, alongside their contexts, in order to

acquire features such as the Dispersion Score, the Combined Similarity Score, the Abstractness Score, and the Bloom's Taxonomy. The thesis is inspired by Bloom's Taxonomy, which is used to measure the quality of the questions and organise them into a hierarchical structure to optimise the learning process (Krathwohl, 2002).

After the features have been obtained, they are then fed into the First SSLP for self-training. Once the training is complete, the independent variables' importance of both the Random Forest Classifier and the Logistic Regression are obtained in order to investigate the most pertinent feature to address the first research sub-question. To answer the question, feature importance analysis revealed that the Combined Similarity Score is the most significant feature of Random Forest Classifier and Logistic Regression (See Figure 4) for the First Semi-Supervised Learning with the Equation 6, while the Bloom's Taxonomy is the most significant feature of Random Forest Classifier and Logistic Regression (See Figure 4c and 4d) for the First Semi-Supervised Learning with the Equation 7.

The second Semi-Supervised Learning relies on the GPT-3 Curie for transfer learning and fine-tuning, taking into consideration the context of the question-answer pairs as well as the pairs themselves.

Both self-training approaches treat correct and incorrect answers identically by design. This decision is made due to the fact that the GPT-3 model used to generate the question-answer pairs is not specifically instructed to create distinguishable differences between correct and incorrect answers. As a result, incorrect answers can be examined explicitly for further analysis.

An approach had been proposed for prompt creation, which entailed the evaluation of labels by Subject Matter Experts using a heuristic approach (Smith et al., 2022). Subsequently, the assessed labels were fed into an Artificial Neural Network to predict a dataset containing unlabeled data. After the predictions, Subject Matter Experts evaluated the outcomes once more in a feedback loop. This strategy is advantageous for the prediction of the unlabeled data in the self-training parts of the SSLP Smith et al., 2022. Despite the benefits, Smith et al., 2022 necessitated human supervision prior to providing the data to the deep learning model. The thesis, however, utilizes GPT-3's token probabilities of the predictions to weakly label the data, thus eliminating the need for human involvement.

Furthermore, no non-automatized human evaluation is employed to evaluate either of the SSLPs during the self-training process. Initially, non-automatized human evaluation can be used to enhance the evaluation process. Subsequently, both of the SSLPs can incorporate the feedback from non-automatized human evaluation to improve their accuracy. Despite GPT-3's impressive outputs suggesting that the evaluation model requires less supervision than state-of-the-art language models (Brown et al., 2020), the lack of human supervision remains a major challenge of this thesis. Crafting an optimal prompt for evaluating question-answer pairs is difficult, as GPT-3 consists of sequence-to-sequence transformers (Brown et al., 2020) and even a single white-space can alter the results. Therefore, the models necessitate a human feedback mechanism to gen-

erate near optimal results.

In order to address the primary research question, two different SSLPs are implemented and their performances are evaluated. The F1 scores of the first SSLP (see Figure 5) with the Equation 6 are presented in Table 3 and 7, while the F1 scores of the same SSLP with the Equation 7 are presented in Table 4 and 8. Additionally, the F1 scores of the second SSLP (See Figure 6) with the Equation 6 are presented in Table 5 and 9, and the F1 scores of the same SSLP with the Equation 7 are presented in Table 6 and 10. As discussed in the literature review, this thesis is motivated by the ability of GPT-3 to perform zero/one/few shot learning (Brown et al., 2020). Furthermore, this thesis takes advantage of GPT-3's pre-trained nature to facilitate transfer learning (Gao et al., 2021), which is especially crucial due to the scarcity of labeled data. This thesis utilises pre-trained models, as demonstrated in Gao et al., 2021, to improve the quality of one/few/zero shot learning with self-training approaches. GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2018) models are one of the primary focuses of this thesis, and both models have been shown to outperform traditional machine learning algorithms in few/zero/one shot learning tasks (Gao et al., 2021). To address the second research sub-question, this thesis leverages pre-trained GPT-3 Curie and BERT models, utilising transfer learning, in order to evaluate the quality of question-answer pairs during the self-training process. Consequently, GPT-3 Curie's self-training approach, which utilizes Random Forest Classifier and Logistic Regression, gives better results than the baseline BERT model in terms of transfer learning.

It is important to consider the preparation of prompts for GPT-3 Curie and GPT-3 Davinci, which respectively generate weakly labeled data and classify Bloom's Taxonomy. Previous research has indicated that the performance of Large Language Models such as GPT-3 is contingent upon the quality of the prompt (Zhou et al., 2022). Consequently, it is essential to recognize that, to attain superior results, the quality of the prompts can be improved either manually or automatically for further research.

The quality of a question-answer pair can be improved with the implementation of multiple rules. This approach is not a heuristic strategy that manipulates the model's final predictions, but rather a process of prompt engineering for GPT-3 models in SSLPs. To achieve this, various rules can be extracted from different teaching resources and prioritized based on their importance. From the experience of this thesis, GPT-3 usually does not perform well with very long prompts. Therefore, the rules must be concise and precise, or multiple prompts must be created with different rules for the models. Both approaches require evaluation, however, the first approach is more advantageous due to performance and cost limitations.

10 Conclusion

This thesis presents the evaluation of question-answer pairs using both GPT-3 Curie and traditional machine learning models such as Random Forest Classifier

and Logistic Regression. To this end, a main research question and two subquestions are examined.

Results from self-training through a minimal amount of data have demonstrated that a fine-tuned GPT-3 model can surpass the performance of traditional machine learning models such as Random Forest Classifiers and Logistic Regression, as well as the baseline model BERT, without any human oversight. Contrary to what has been reported in prior research, the two distinct SSLPs do not require any heuristic intervention or human supervision of the predictions. The two proposed weakly labelling strategies play a critical role in initiating Semi-Supervised Learning without relying on manual labelling. This implies that there is no further requirement for human input throughout both pipelines, thereby conserving time, energy, and resources. GPT-3 Curie can facilitate the quality assessment of question-answer pairs without human intervention by employing an Semi-Supervised Learning feedback loop with their tokens' conditional probabilities. Retrieving the token probabilities can be accomplished in two ways: first, by summing the token probabilities of each sentence and fitting them into a Softmax with Equation 6, or by taking the average of the token probabilities of each sentence with Equation 7. These strategies indicate the confidence levels of each weak label generated by GPT-3. To determine the best feature for Semi-Supervised Learning to reinforce weakly labeled data for assessing questions and answers, feature importance from Scikit-learn (Pedregosa et al., 2011) using permutation_importance function is utilized. Additionally, two different Semi-Supervised Learning strategies are employed (see Equations 6 and 7). Subsequently, according to the feature importance analysis, the Combined Similarity Score and the Bloom's Taxonomy are the most significant independent variables for the First Semi-Supervised Learning with Equation 6 and 7 for both Random Forest Classifier and Logistic Regression, respectively. Consequently, the thesis can reduce the time and cost associated with manual labeling due to its self-training approach. Furthermore, it can be beneficial for the EdTech industry to more accurately evaluate student performance. As a result, EdTech companies can reallocate their budget towards more innovative and research-oriented strategies, rather than relying heavily on non-automatized human evaluation. This thesis can be of great value to scholars who are striving to enhance the quality of text assessment. The results of this thesis have indicated that two distinct SSLPs have achieved considerable outcomes. However, certain aspects should be considered for potential future research. The Combined Similarity Score and The Bloom's Taxonomy, being the two of the most significant independent variables, are based on the data that being worked with. Thus, as more data is added and improvements are made, the significance of the Combined Similarity Score and the Bloom's Taxonomy may vary. Additionally, the evaluation of the results is not conducted by human labour, which remains a major challenge for the thesis. Furthermore, the evaluation model does not distinguish between correct and incorrect answers, as the thesis does not evaluate the correctness of the question-answer pairs. To overcome this, different prompts or pipelines can be utilized for correct and incorrect answers respectively. Furthermore, the prompts of GPT-3 Curie, which

is responsible for creating weakly labeled data and self-training in the Second SSLP, and GPT-3 Davinci, which is responsible for classifying the Bloom's Taxonomy of the given questions, can be improved to produce better evaluation results.

11 Acknowledgements

I would like to thank my Supervisor dr. Çiçek Güven, Second Reader dr. Gonzalo Nápoles, External Supervisor Zülküf Genç, and my friends for their support for me and my thesis.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman,

Boecking, B., Neiswanger, W., Xing, E., & Dubrawski, A. (2021). Interactive weak supervision: Learning useful heuristics for data labeling. *International Conference on Learning Representations*. https://openreview.net/forum?id=IDFQI9OY6K

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. https://arxiv.org/abs/2005.14165

Chapelle, O., Scholkopf, B., & Zien, A., Eds. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542–542. https://doi.org/10.1109/TNN.2009.2015974

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

Hernandez, A. O., Sreedharan, S., & Kambhampati, S. (2021). Gpt3-to-plan: Extracting plans from text using GPT-3. *CoRR*, *abs/2106.07131*. https://arxiv.org/abs/2106.07131

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Meng, Y., Shen, J., Zhang, C., & Han, J. (2018). Weakly-supervised neural text classification. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 983–992. https://doi.org/10.1145/3269206.3271737

Nashaat, M., Ghosh, A., Miller, J., Quader, S., Marston, C., & Puget, J.-F. (2018). Hybridization of active learning and data programming for labeling large industrial datasets. *2018 IEEE International Conference on Big Data (Big Data)*, 46–55. https://doi.org/10.1109/BigData.2018.8622459

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, *39*(2), 103–134.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Shihadeh, J., Ackerman, M., Troske, A., Lawson, N., & Gonzalez, E. (2022). Brilliance bias in gpt-3. *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, 62–69. https://doi.org/10.1109/GHTC55712.2022.9910995

Skillsoft. (2022). About us. https://www.skillsoft.com/about

Smith, R., Fries, J. A., Hancock, B., & Bach, S. H. (2022). Language models in the loop: Incorporating prompting into weak supervision. https://doi.org/10.48550/ARXIV.2205.02318

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Zhong, R., Snell, C., Klein, D., & Steinhardt, J. (2022). Describing differences between text distributions with natural language. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 27099–27116). PMLR. https://proceedings.mlr.press/v162/zhong22a.html

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. https://doi.org/10.48550/ARXIV.2211.01910

Zhu, X. J. (2005). Semi-supervised learning literature survey.

12 Appendix

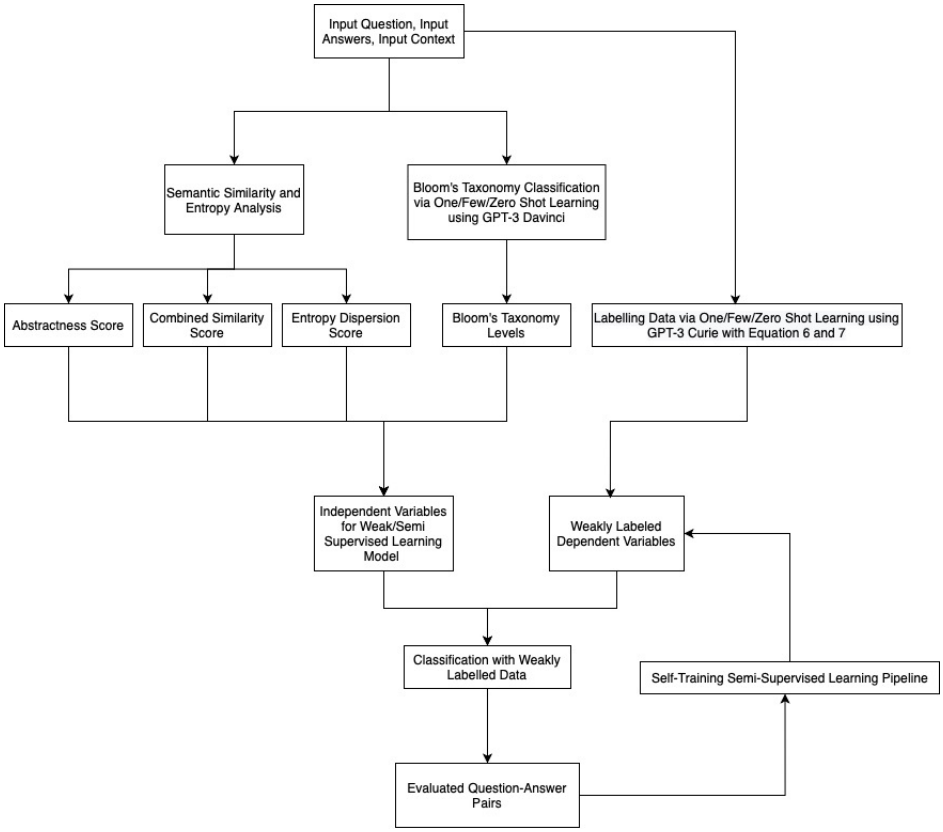


Figure 5: SSLP and Feature Generation

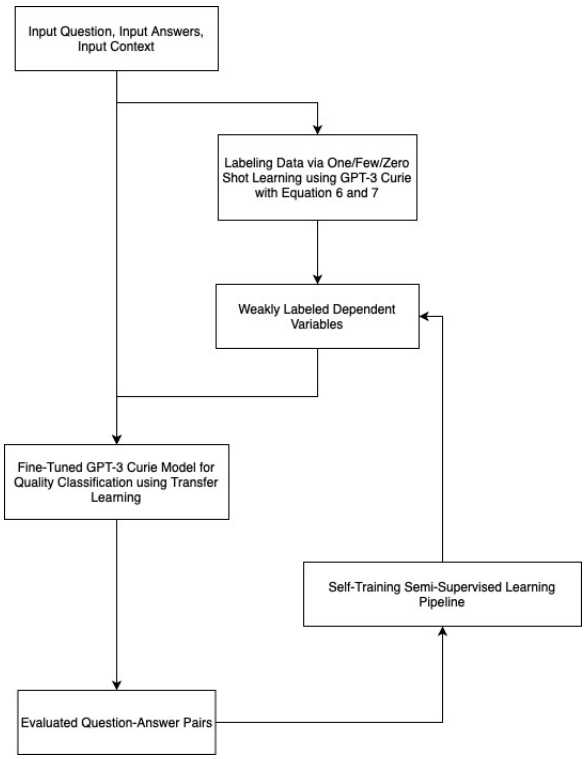


Figure 6: Second SSLP and Feature Generation