International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2023, Vol. 19, Issue 3, pp. 259-278

Revolutionising Educational Assessment: Automated Question Classification using Bloom's Taxonomy and Deep Learning Techniques – A Case Study on Undergraduate Examination Questions

Kuhaneswaran Banujan, Samantha Kumara, Senthan Prasanth & Nirubikaa Ravikumar
Sabaragamuwa University of Sri Lanka

ABSTRACT

Examinations are one way of evaluating students. To ensure the production of valid exams, frameworks such as Bloom's taxonomy are utilised when preparing questions. Bloom's taxonomy is a well-known framework that categorises educational objectives into six hierarchical levels of cognitive complexity. However, manually categorising exam questions can be time-consuming and subjective. The extant literature has yet to leverage advanced deep learning methods and state-of-the-art word embedding techniques. This study utilises the effectiveness of Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) models along with GloVe, BERT and TF-IDF for automating the classification of exam questions according to the revised Bloom's taxonomy. The study collected various question types from online sources and multiple state universities in Sri Lanka, resulting in a dataset of 16,584 questions labelled manually with the aid of domain experts. The dataset was cleaned using natural language processing techniques. Three models were proposed: ANN+TF-IDF, LSTM+GloVe, and LSTM+BERT. The results of the ANOVA and post hoc pairwise comparisons using Bonferroni correction indicate that the LSTM+BERT model outperformed the other models significantly. The proposed approach provides a reliable and consistent way of evaluating students, and educators can use it to improve their teaching strategies. The findings of this study have important implications for educational institutions and can lead to more effective and efficient evaluations.

Keywords: E*xaminations, Bloom's taxonomy, deep learning, ANN, LSTM, GloVe, BERT*

INTRODUCTION

In educational institutions, the traditional and conventional method of evaluating students is through written examinations (Mohammed & Omar, 2020; Zhang et al., 2021). However, preparing appropriate exam questions to achieve the desired course outcomes can be challenging for examiners (Jayakodi et al., 2016b). To ensure the production of high-quality exams, many lecturers follow frameworks such as Bloom's taxonomy or revised Bloom's taxonomy while preparing exam questions (Mohammed & Omar, 2018). Bloom's taxonomy (Bloom, 1956) or revised Bloom's taxonomy (Anderson et al., 2001) is a well-known framework which comprises three learning domains:

- the Cognitive domain: primarily concerned with intellectual abilities including critical thinking, problem solving, and knowledge building.
- the Affective domain: concerned with leaners attitudes, values, interests, and appreciation, and
- the Psychomotor domain: includes students' physical task-accomplishment, mobility, and skill-performance abilities.

The taxonomy has been widely used in educational settings to develop instructional objectives, design curricula, and evaluate learning outcomes. However, some academicians lack knowledge of Bloom's taxonomy or revised Bloom's taxonomy (Omar et al., 2012), and some were unable to

distinguish the difference between its various levels, leading to misclassification and poor quality examinations (Omar et al., 2012, Jayakodi et al., 2016a).

Additionally, an exam question often falls under many assessment categories within a particular taxonomy. Hence, it is challenging to classify exam questions and determine the section of each taxonomy level assessment to which they belong (Jayakodi et al., 2016b). Hence, students should be able to recall, express and apply what they have learned to new and challenging outcomes. As an alternative to conventional approaches, question paper writers must utilise Bloom's taxonomy or revised Bloom's taxonomy principles when assessing pupils holistically. However, according to Bloom's taxonomy or revised Bloom's taxonomy, assessing question papers for their specificity and complexity may be time-consuming (Jain et al., 2019).

Classifying exam questions according to Bloom's taxonomy or revised Bloom's taxonomy can help educators evaluate the effectiveness of their teaching strategies and identify areas for improvement. Classifying questions according to Bloom's Taxonomy Cognitive Domain (BTCD) provides several benefits, including an appropriate and effective way to measure students' intellectual abilities (Jain et al., 2019) and covering a range of thinking skills that range from the simplest to the most complex (Mohammed and Omar, 2018). Furthermore, automated classification of exam questions according to Bloom's taxonomy or revised Bloom's taxonomy can benefit both teachers and students. It can save educators time and effort in evaluating exams, allowing them to concentrate on other aspects of instruction. In addition, it can offer insights into the efficacy of teaching strategies and pinpoint areas for improvement. It can give students a more objective and accurate evaluation of their learning outcomes, enabling them to identify their strengths and limitations and modify their study practices accordingly (Yahya and Osman, 2011). However, manually categorising exam questions can be time-consuming and subjective, making it difficult to achieve consistency and reliability in the evaluation process.

Automated approaches to question classification utilising machine learning techniques have demonstrated the potential to accelerate the categorisation of each assessment question, thereby reducing the amount of additional effort required to integrate external teaching repositories (Zhang et al., 2021). Data mining is a process that uses statistical methods, mathematics, artificial intelligence, and machine learning to extract meaningful data and knowledge from massive datasets (Han et al., 2012). On the other hand, text mining extracts information from a collection of documents using analytical methods, such as classification, which is one of the components of data mining (Aninditya et al., 2019).

Question classification differs from document classification, as questions are typically written in short forms. Unlike document classification, which benefits from extensive available information, short texts often lack context and sparsity (Yang et al., 2013, Wang et al., 2016). Consequently, pure statistical methods such as N-gram and TF-IDF are unsuitable for question classification because these methods require vast amounts of data to achieve high accuracy (Abduljabbar and Omar, 2015). Additionally, question classification, especially regarding the cognitive domain, differs from general text classification tasks focusing on topic classification. Furthermore, while general text classification tasks typically use formal language and strict grammar, teachers often use colloquial language and write in short forms when teaching in class. Consequently, it is challenging to classify questions with one sentence, even after preprocessing them with several words (Huang et al., 2021). However, accurately classifying exam questions according to Bloom's taxonomy or revised Bloom's taxonomy can be challenging.

Machine learning and deep learning techniques such as support vector machine (SVM), k-nearest neighbour (k-NN), naive Bayes (NB), decision tree (DT), random forest (RF), artificial neural networks (ANN) (McCulloch and Pitts, 1943) and long short-term memory (LSTM) (Schmidhuber and Hochreiter, 1997) can help address these challenges by analysing the structure and content

of exam questions and learning patterns indicative of each taxonomy level. Through the process of instructing these algorithms with extensive sets of examination queries that have been classified based on Bloom's taxonomy or revised Bloom's taxonomy, they can acquire the ability to identify fundamental patterns and effectively categorise novel questions with a high degree of precision. Word embedding is an important step in natural language processing (NLP) because it permits the representation of words in a numerical format that is readily handled by machine learning algorithms. Word embedding techniques such as Word2Vec, Bag of Words, Global Vectors (GloVe) (Pennington et al., 2014), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), and Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck Jones, 1972) can also improve the accuracy of the classification process by capturing the semantic and contextual relationships between words in exam questions.

By integrating machine learning and word embedding techniques, attaining a heightened level of precision in categorising examination queries based on either Bloom's taxonomy or revised Bloom's taxonomy becomes feasible. However, it is important to note that the quality of the classification process depends on the training data quality and the machine learning model design. Therefore, careful consideration must be given to the selection of the training dataset and the parameters of the machine learning models to ensure the accuracy and reliability of the classification process. In summary, the automated classification of exam questions according to the revised Bloom's taxonomy using machine learning and word embedding techniques has the potential to improve the efficiency and accuracy of educational assessment practices. By leveraging the power of these techniques, teachers can obtain more objective and reliable evaluations of students' learning outcomes, and students can receive more personalised feedback on their strengths and weaknesses.

SIGNIFICANCE OF THE STUDY

The study aims to contribute to this field by exploring the effectiveness of ANN and LSTM along with GloVe, BERT, and TF-IDF for classifying exam questions into each revised Bloom's taxonomy category. To the best of our knowledge, this is the first attempt to apply deep learning to classify educational objectives in revised Bloom's levels.

The key contributions of the proposed research are as follows:

- Incorporation of multiple word embedding techniques: Although some prior research employed single word embedding approaches such as Word2Vec or TF-IDF, the proposed methodology uses several word embedding techniques such as GloVe, BERT, and TF-IDF to capture the semantic and contextual relationship between words. By merging different word embedding approaches, it is expected that the classification process would become more accurate and robust.

- Use of ANN and LSTM: While many previous studies have employed machine learning techniques such as SVM, k-NN, NB, DT, RF, and MLP, the proposed approach employs deep learning techniques such as ANN and LSTM, which are more powerful and flexible algorithms for handling sequential data such as text. Using these deep learning approaches, will identify more complicated patterns and relationships in exam questions.

- Evaluation on a larger dataset: Although some earlier studies have employed limited datasets of a few hundred exam questions, our proposed approach uses a larger dataset to obtain more accuracy and reliability in assessing students' learning outcomes.

The proposed study has the potential to considerably enhance the efficacy and accuracy of educational assessment processes and provide valuable insights into students' learning outcomes.

The results of this research may be used in various educational environments, such as e-learning platforms, to give more tailored and objective assessments of students' learning outcomes and improve education quality.

Here are some examples of how the research can be applied in various educational settings:

- E-learning platforms: Automatic categorisation of exam questions according to the revised Bloom's taxonomy may be included in e-learning platforms to give more efficient and accurate assessments of students' learning results. By utilising the capabilities of deep learning and word embedding methods, the suggested method can effectively categorise examination questions and offer students individualised feedback on their strengths and shortcomings. This may assist pupils in modifying their study habits and fostering higher-order thinking abilities.
- Lecturer assessment: Lecturers may save time and effort reviewing tests using the suggested method, enabling them to concentrate on other elements of instruction. The automated categorisation of examination questions may give objective and consistent assessments of student learning results and pinpoint areas where teaching practises might be improved.
- Educational research: The suggested method may also be used in educational research to evaluate the efficiency of instructional tactics and find areas for improvement. By assessing the cognitive difficulty of examination questions and the learning goals of each updated category of Bloom's taxonomy, researchers may acquire valuable insights about students' cognitive development and the efficacy of teaching practices.
- Standardised testing: The suggested method may also be used to verify the consistency and reliability of exam assessments in standardised testing. By automating the categorisation of exam questions according to the revised Bloom's taxonomy, standardised testing firms may verify that each exam question is graded consistently and adequately and encourage test-takers' higher-order thinking abilities.

Overall, the research has a wide range of applications in various educational settings, from e-learning platforms to standardised testing, and can contribute to ongoing efforts to improve educational assessment practices and promote higher-order thinking skills in students.

LITERATURE REVIEW AND RELATED WORK

Bloom's taxonomy is a framework developed by Benjamin Bloom (Bloom, 1956) to describe levels of learning objectives that promote higher-order thinking skills. It consists of six levels, with each level building upon the previous one: knowledge, comprehension, application, analysis, synthesis, and evaluation. In 2001, Lorin Anderson and David Krathwohl (Anderson et al., 2001) revised Bloom's taxonomy to better reflect the changing needs of education in the 21st century. The revised taxonomy has six levels, with the names changed to better represent the cognitive processes involved: Remember: retrieving pertinent information from long-term memory; Understand: constructing meaning from instructional messages, including oral, written, and visual communication; Apply: using a process via execution or implementation; Analyse: separating material into its component pieces, identifying how those parts relate to one another and a larger structure or goal; Evaluate: evaluating based on criteria and standards; and Create: bringing together pieces to make a cohesive or functioning whole; rearranging elements to produce a new pattern or structure.

One of the main differences between the original and revised taxonomy is the emphasis on the verbs used to describe the cognitive processes involved. The revised taxonomy uses more action-oriented verbs that better reflect the learning process (Forehand, 2005). The revised Bloom's taxonomy provides a more comprehensive and relevant framework for educators to design learning

objectives and assessments. It emphasises the skills needed for success in the 21st century, such as problem solving, critical thinking, and creativity. Figure 1 shows the comparison between Bloom's taxonomy and the revised Bloom's taxonomy (Santos et al., 2021).

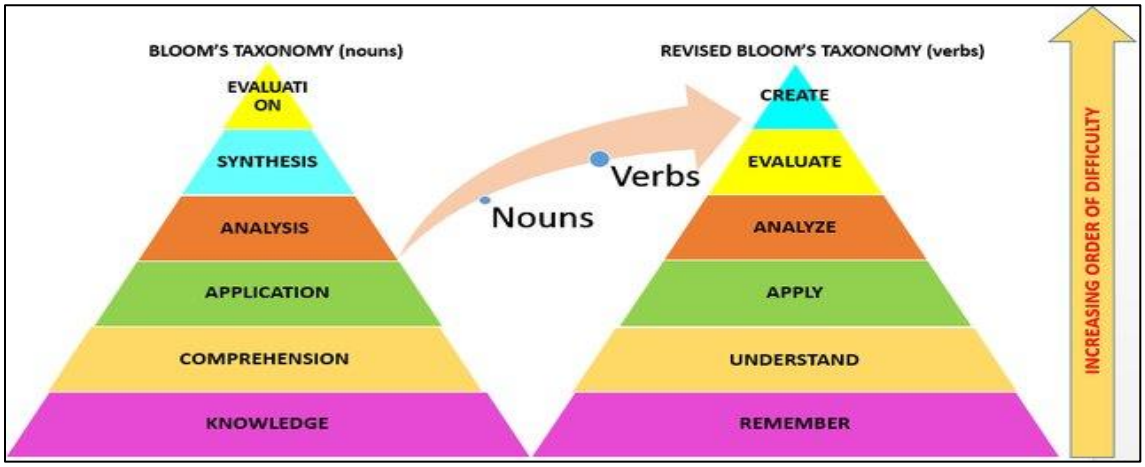

Figure 1: Comparison between Bloom's taxonomy and revised Bloom's taxonomy (adopted from Santos et al., 2021).

Zhang, Wong, Giacaman, & Luxton-Reilly (2021) proposed a deep learning pipeline for improved question classification into Bloom's taxonomy domains. The proposed approach employs preprocessing techniques, including stop-word removal, stemming, and tokenization, and advanced machine learning techniques, such as convolutional neural networks (CNNs) and LSTM networks. The approach was tested on a dataset of over 16,000 questions and accurately classified questions into the six domains of Bloom's taxonomy. The study concluded that the proposed approach can be useful in developing better assessments covering all of Bloom's taxonomy levels and improving the quality of educational assessments. The authors of "Automated Classification of Computing Education Questions using Bloom's Taxonomy" used Google's BERT to create a machine-learning technique for categorising programming questions based on Bloom's taxonomy. The Canterbury Question Bank, which professionals in computer education classified, was the source of questions. The findings demonstrated that the model could predict categories with reasonable accuracy but performed better when classifying questions at lower levels of Bloom's taxonomy. The research demonstrated the potential for machine learning to aid instructors in assessing assessment items (Zhang et al., 2021).

Mohammed & Omar (2020) offered a classification approach for automatically categorising test questions based on Bloom's taxonomy across academic areas. The suggested approach for classifying questions comprises extracting two features: TFPOS-IDF and word2vec. The TFPOS-IDF function computes the term frequency-inverse document frequency depending on the part of speech to provide appropriate weights to key terms in the inquiry. In contrast, the word2vec feature employs word2vec, which has already been trained to enhance the categorisation process. These attributes are supplied to three distinct classifiers to categorise the questions: k-NN, Logic Regression (LR), and SVM. The research used two datasets, one including 141 questions and the other containing 600 questions. Their findings show the efficiency of the suggested strategy in categorising questions from numerous domains based on Bloom's taxonomy.

In the paper titled "Exam Questions Classification Based on Bloom's Taxonomy Cognitive Level Using Classifiers Combination," Abduljabbar & Omar, (2015) suggested a novel automated approach for categorising exam questions according to the cognitive levels of Bloom's taxonomy.

The suggested technique employed a combination strategy based on a voting algorithm that combined SVM, NB, and k-NN classifiers. To identify questions as having or lacking feature selection, Chi-Square, Mutual Information, and Odd Ratio were also regarded as feature selection approaches. The combination algorithm was used to combine the total performance of the three classifiers, and the mutual information feature selection approach offered the maximum classification accuracy. This research aimed to improve the classification procedure's accuracy by combining various feature selection techniques and classification algorithms. The results suggest that the proposed strategy is promising and comparable to other models with similar characteristics (Abduljabbar & Omar, 2015). A study by Huang et al. (2021) aimed to enhance the effectiveness of the curriculum design process for teachers by utilising machine learning to classify their questions automatically. The methodology employed in this study entailed generating keywords and extracting TF-IDF features for question classification. The research findings indicated that the utilisation of personalised keywords is a critical factor in attaining a considerable degree of precision in categorising questions based on Bloom's taxonomy. This approach resulted in an accuracy level of 86.0%. The authors concluded that this approach improved performance and reduced the number of features required for classification, making it applicable across various subjects.

The primary objective of the work titled "WordNet and Cosine Similarity-based Classifier of Exam Questions Using Bloom's Taxonomy" was to automatically classify exam questions according to their learning levels using Bloom's taxonomy. Jayakodi et al., 2016b applied natural language processing (NLP) methods such as tokenisation, stop word removal, lemmatisation, and tagging were prior to classification. Next, using NLTK and cosine similarity algorithms, WordNet similarity algorithms were created to build a unique set of criteria for determining the question category and weight. Exam questions may be readily assessed using this method, and exam papers can be redesigned depending on the conclusion of this categorisation procedure. The evaluation was based on a sample of examination questions from the Department of Computing and Information Systems at Wayamba University in Sri Lanka. The assignment of weights was determined by the total value produced by the WordNet and cosine algorithms. A domain expert validated question category identification (Jayakodi et al., 2016b).

In the study by Mohammed and Omar (2018), the authors utilised a dataset of 600 questions evenly distributed across each level of Bloom's taxonomy. The questions were subjected to a preprocessing phase to ensure their suitability for implementing the proposed enhanced feature. Three machine learning classifiers were used for classification: SVM, NB, and k-NN. The enhanced feature yielded satisfactory results, outperforming the traditional feature TF-IDF across all classifiers in terms of weighted recall, precision, and F1-measure. SVM emerged as the top-performing classifier. Aninditya et al. (2019) assessed the effectiveness of the NB classifier in classifying questions based on the cognitive level of Bloom's taxonomy. To achieve this, the authors used a real-world dataset of mid-term and final exam questions from the Department of Information Systems at Telkom University for the academic years 2012/2013 to 2018/2019. The study examined various indexing terms, such as Words, Characters, and N-grams.

The authors of the paper "Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings" offer a deep learning model based on LSTM to classify course learning outcomes (CLOs) and assessment items based on various levels of Bloom's cognitive domain. Initially, Shaikh et al., (2021) classified CLOs and evaluation items into Bloom's taxonomy with an overall accuracy of 55% using a keyword-based approach to their datasets. The proposed model predicted Bloom's level for CLO and assessment question items. The suggested model has a simpler architecture than previous deep learning models published in the literature. It obtained a classification accuracy of 87% for CLOs and 74% for assessment question items. Compared to a previous study for the same task, the proposed model demonstrates a 3% improvement in overall accuracy (Shaikh et al., 2021). Table 1 presents a comparison between the approach proposed in this paper and existing studies.

Table 1: *Comparison Between the Proposed Research and Existing Studies*

Study	Dataset	Domain	Algorithms Used	Feature Selection Methods	Blooms or Revised
Kusuma et al. (2015)	130	Common	SVM	Lexical feature extraction	Revised
Ifham et al. (2022)	141	Common	SVM, ANN	Word Embedding (TF-IDF)	Blooms
Supriyanto et al. (2013)	274	Common	NB	Chi-Square and Information Gain	Blooms
Aninditya et al. (2019)	300	Information System	NB	Word Embedding (TF-IDF)	Revised
Zhang et al. (2021)	504	Computing	BERT	Word Embedding (BERT)	Blooms
Mohammed and Omar (2018)	600	Common	SVM, NB, k-NN	Word Embedding (TF-IDF)	Blooms
Osman and Yahya (2016)	600	English Language	NB, LR, SVM, DT	Word Embedding (Bag of Words and n-grams)	Blooms
Yahya et al. (2012)	600	Common	SVM	Word Embedding (TF-IDF)	Blooms
Huang et al. (2021)	1,000	Common	LR, RF, XGBoost	Word Embedding (TF-IDF)	Blooms
Abduljabbar and Omar (2015)	Not mentioned	Computer Programming	SVM, NB, k-NN	Chi-Square, Mutual Information	Blooms
Proposed Research	15,000	Common	LSTM, ANN	Word Embedding (GloVe, TF-IDF)	Revised

Compared to existing studies, the proposed approach has several novel aspects, namely:

- Incorporation of multiple word embedding techniques: While some previous studies have used single-word embedding techniques such as Word2Vec, the proposed approach incorporates multiple word embedding techniques such as GloVe, BERT, and TF-IDF to capture the semantic and contextual relationships between words. The intention is to enhance the accuracy and robustness of the classification process by combining various word embedding techniques;
- Use of ANN and LSTM: While many previous studies have used machine learning techniques such as SVM, k-NN, NB, DT, RF, and MLP, the proposed approach uses deep learning techniques such as ANN and LSTM, which are more powerful and flexible algorithms for handling sequential data such as text. The aim is to capture more complex patterns and dependencies in exam questions by utilising these deep learning techniques;
- Evaluation on a larger dataset: While some previous studies have used small datasets of a few hundred exam questions, our proposed approach aims to utilise a larger dataset to achieve higher accuracy and reliability in evaluating students' learning outcomes.

To summarise, the approach proposed here involves the integration of various word embedding techniques, applying deep learning techniques such as ANN and LSTM, utilising a larger dataset, and implementing a comprehensive evaluation methodology. These elements collectively offer a new and more resilient approach to the classification of exam questions based on the revised Bloom's taxonomy.

METHODOLOGY

Figure 2 below depicts the high-level methodological framework and the detailed steps carried out during this research.

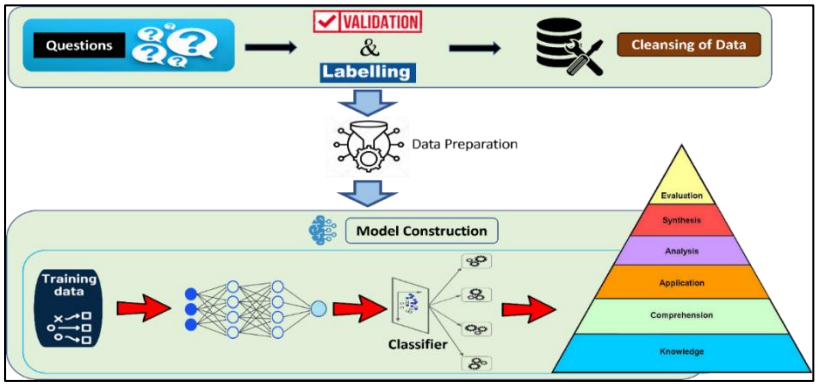

Figure 2: Proposed Methodological Framework

Data Collection

The study collected various question types from online sources and multiple state universities in Sri Lanka, resulting in a dataset of 16,584 questions that were carefully analysed. The revised Bloom's taxonomy classification system includes six categories: remember, understand, apply, analyse, evaluate and create. Each category in the dataset consists of approximately 2,500 questions.

Data Validation and Labelling

A team of 8 members, including domain experts and the authors, individually reviewed each question to assess its appropriateness concerning the revised Bloom's taxonomy categories. Any questions that were irrelevant to any of the categories were removed from the dataset. As a result, 1,584 questions out of the original 16,584 were identified as nonrelevant to any of the revised Bloom's taxonomy categories and were subsequently removed. Once the irrelevant questions were removed, the appropriate labels corresponding to each question's revised Bloom's taxonomy level were manually added. Table 2 below provides examples of questions in the dataset and their corresponding labels.

Table 2: *Sample Questions and Their Respective Labels*

Sample Questions	Taxonomy
Define substrate, floodplain, and streambank.	Remember
Describe the typical computer peripheral parts.	
Give a detailed explanation of how to copy text from one application to another.	Understand
Give an example of statistical software quality assurance using the Pareto Principle.	
Determine a beam's deflection under uniform loading.	Apply
How could someone modify the Cornell Method to make it effective for researching a topic for a paper?	
Separate macroeconomics from microeconomics.	Analyse
How do psychologists differ in their general attitudes toward third-party presence?	

Give examples of how the biological idea of symbiotic relationships may be applied to help resolve socially induced issues such as water pollution, overflowing landfills, or homelessness.	Evaluate
Imagine yourself recommending a book to someone as a librarian. Write a paragraph outlining your position.	
Considering the fundamental kinetic knowledge and determining the degree of galvanic connection between two metals.	Create
Using the healthy eating guidelines, create a menu you believe most people will appreciate.	

Cleansing of Data

After the labelling process, the entire question set underwent various steps to prepare it for further analysis. As a prerequisite, unnecessary terms and phrases were eliminated, and several data cleansing methods were applied, as shown in Figure 3. The following are some of the significant techniques used during this process: removal of stop words, symbols, and expressions, tokenization, removal of extra whitespaces, and stemming.

The first step involved removing stop words and converting all questions to lowercase. Removing stop words helped maintain consistency in the dataset and improve the model's accuracy. Punctuations, expressions, and quotes were removed from the textual data to treat each question equally.

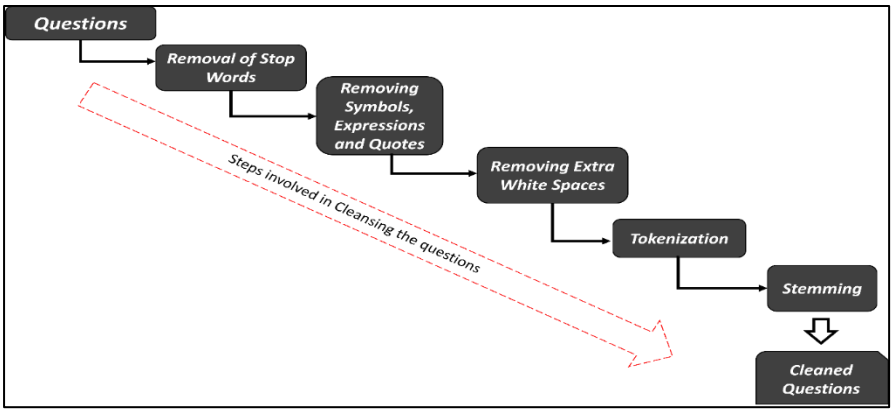

Figure 3: The processes involved in cleaning the questions

Extra white spaces from the questions were removed to obtain responses promptly and reduce the workload during the analysis. Tokenization was used to divide the original text into manageable pieces, such as words and sentences, to aid context comprehension or model development for NLP. Stemming was also applied to lower inflection towards their root forms. After using these processes for each question, the dataset is represented in Table 3 below.

Table 3: *Questions and Their Representation Once After Each Process Is Carried Out*

Question	Stop words removal & converting to lowercase	Tokenisation	Stemming
How could someone modify the Cornell Method to make it effective for researching a topic for a paper?	'could', 'someone', 'modify', 'cornell', 'method', 'make', 'effective', 'researching', 'topic', 'paper'	'how', 'could', 'someone', 'modify', 'the', 'cornell', 'method', 'to', 'make', 'it', 'effective', 'for', 'researching', 'a', 'topic', 'for', 'a', 'paper', '?'	'could', 'someone', 'modify', 'cornell', 'method', 'make', 'effective', 'researching', 'topic', 'paper'
How do psychologists differ in their general attitudes toward third-party presence?	psychologists', 'differ', 'general', 'attitudes', 'toward', 'thirdparty', 'presence'	'how', 'do', 'psychologists', 'differ', 'in', 'their', 'general', 'attitudes', 'toward', 'third-party', 'presence', '?'	'psychologist', 'differ', 'general', 'attitude', 'toward', 'thirdparty', 'presence'

Data Preparation

After cleansing the questions, they were converted into vectors to generate feature vectors for training machine learning algorithms. Using the Transformer architecture, pretrained word contextualised embedding methods were used to form the vectors. Word embedding represents words with numerical representations in an N-dimensional dense vector with identical meanings. Pretrained word embedding techniques such as word2vec, GloVe, and TF-IDF can be used, or an embedding model can be trained using a provided corpus.

This research utilised contextualised embedding methods such as BERT, TF-IDF, and GloVe to turn questions into numerical vectors to determine the technique combination that produced the maximum classification accuracy. BERT employs a transformer, an attentional system that identifies word associations in a text. GloVe, on the other hand, is a multidimensional vector that illustrates how a word links to other words. At the same time, TF-IDF assigns each word a single value devoid of semantic importance. Several strategies for word embedding were examined, including a counter vectorizer, a bag of words, Word2Vec, and one-hot encoding. The encoder mechanism of BERT was used to construct a language model, and combinations of BERT and LSTM, GloVe and LSTM, and TF-IDF and ANN were evaluated to determine the most effective pair approaches.

Construction of the models

During this study, LSTM+BERT, LSTM+GloVe, and ANN+TF-IDF were implemented and compared against each other for the highest classification accuracy in identifying the question levels in the revised Bloom's taxonomy.

Implementing the ANN Model

ANN is also a computational model replicating nerve cell behaviour in the human brain. A typical ANN model consists of three layers: the input, hidden, and output layers. Table 4 represents the configuration parameters discovered for the ANN with TF-IDF to obtain the optimum results.

Table 4: *Parameters Configured with ANN*

Parameter	Value for TF-IDF
Epochs	25
Batch size	32
Optimiser	Adam
Loss	categorical_crossentropy
Activation (first dense layer)	ReLu
Activation (output layer)	SoftMax

Implementing the LSTM Model with BERT and GloVe

LSTMs are recurrent neural network (RNN) types that can learn long-term dependencies. They are commonly used and have shown excellent performance in various tasks. LSTMs are specifically designed to avoid the issue of long-term dependency, and their default behaviour is to remember information for extended periods rather than to strive for learning. Table 5 shows the configuration parameters for the LSTM-based models.

Table 5: *Parameters Configured with LSTM*

Parameter	Value for GloVe	Value for BERT
Epochs	25	25
Batch size	32	32
Optimiser	Adam	Adam (learning_rate=2e-5)
Loss	categorical_crossentropy	categorical_crossentropy
Activation (first dense layer)	ReLu	ReLu
Activation (output layer)	SoftMax	SoftMax

Model Evaluation

In NLP, various evaluation matrices are used to assess the performance of a model. These matrices measure different aspects of the model's accuracy and effectiveness. This article will discuss NLP's most commonly used evaluation matrices and their relevance to the proposed model.

Accuracy: Accuracy (shown as Equation 1 below) is a commonly used evaluation metric in NLP. It measures the percentage of the total number of correctly classified predictions.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} * 100 \quad (1)$$

Precision: Precision (shown as Equation 2 below) is another commonly used evaluation metric in NLP. It measures the percentage of correct positive predictions out of all positive predictions.

$$Precision\ = \frac{True\ Positive}{True\ Positive + False\ Positive} * 100 \quad (2)$$

Recall: Recall (shown as Equation 3 below) is a metric measuring the percentage of correctly predicted positive instances out of the total positive samples.

$$Recall\ = \frac{True\ Positive}{Total\ Number\ of\ Predictions} * 100 \quad (3)$$

F1 Score: The F1 Score (shown as Equation 4 below) is a metric that is a combination of precision and recall.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Loss Function: The loss function (shown as Equation 5 below) measures the expected and actual output deviation, where *y* is the actual probability distribution and *p* is the predicted probability distribution.

$$Loss\ Function\ = \sum ylog(p) \quad (5)$$

Mean Absolute Error (MAE): MAE (shown as Equation 6 below) is a metric that measures the average absolute difference between the predicted values and the actual values, where n is the total number of samples, y is the actual value, and y_pred is the predicted value.

$$MAE = \left(\frac{1}{n}\right) * \sum|y - y\_pred| \quad (6)$$

Mean Squared Error (MSE): MSE (shown as Equation 7 below) measures the average squared difference between the predicted and actual values, where n is the number of data points, y is the actual value, and $\hat{y}$ is the predicted value.

$$MSE = \left(\frac{1}{n}\right) * \sum(y - \hat{y})^2 \quad (7)$$

Several evaluation matrices, including accuracy, precision, recall, F1 score, loss function, MAE, MSE, and AUC-ROC, will be used to assess the proposed model. These matrices give several indices of the model's accuracy and efficacy, enabling a thorough assessment of the model's performance in NLP tasks.

RESULTS AND DISCUSSION

Results obtained from the LSTM + BERT approach

Assessment of a neural network (NN) model often entails measuring the error value using a loss function. Training an NN entails fine-tuning the weights and biases derived from input data to minimise the loss function and provide outcomes with minimum error. The results are shown in the Figure 4 below as follows:

- Figure 4(a) shows the accuracy curve for both the training and testing data segments, which reveals that the final model obtained an accuracy of 88.7% on the testing data.
- Figure 4(b) shows the curve acquired for the loss function for the LSTM+BERT combined model, which produced a loss function value of 1.25 for the training partition.
- Figures 4(c) and (d) show the graphs of MAE and MSE against epochs, respectively.
- Figures 4(e) and 4(f) show the model's precision and recall values for both training and testing data.

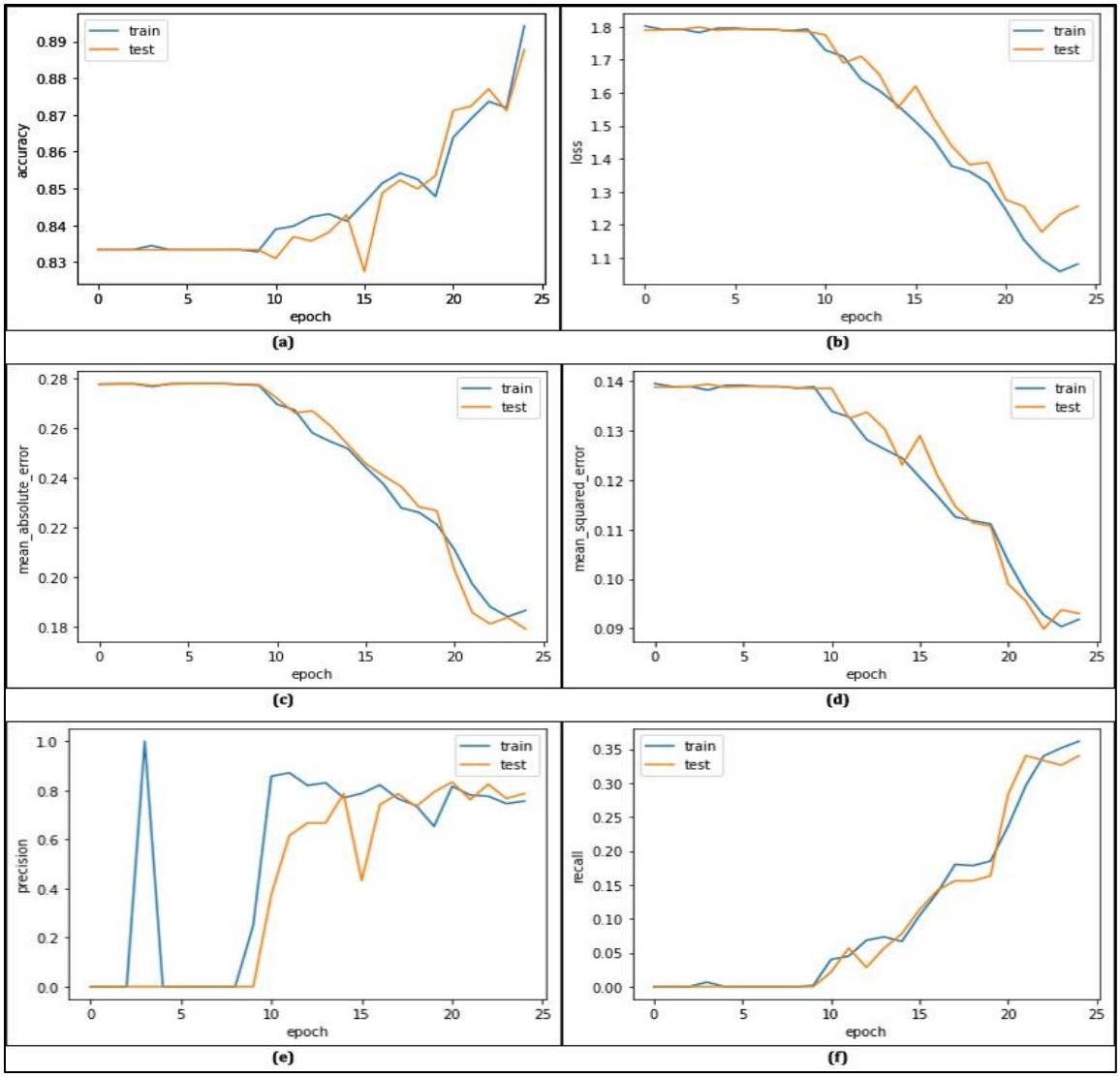

Figure 4: (a) Accuracy, (b) loss function, (c) MAE, (d) MSE, (c) precision, and (f) recall for the LSTM+BERT model

The true positive rate (TPR) is displayed on the y-axis, and the false positive rate (FPR) is plotted on the x-axis of the AUC-ROC graph. The optimal model would have TPR = 1 and FPR = 0 for an AUC-ROC score of 1. The AUC-ROC score may be used to evaluate the performance of various models. A higher AUC-ROC score suggests superior performance since the model has a greater TPR and lower FPR. In addition, the shape of the curve might offer insight into the model's performance. If the curve is closer to the graph's upper-left corner, the model can differentiate between positive and negative examples more. The model's performance is comparable to random guessing if the curve is closer to the diagonal. Overall, the AUC-ROC graph visually represents the model's performance and can be used to determine the optimal threshold for classification.

Figure 5 shows the AUC-ROC curve for (a) Remember, (b) Understand, (c) Apply, (d) Analyse, (e) Evaluate and (f) Create for the LSTM+BERT model. As per the illustrations, the LSTM + BERT

model performs well in predicting Remember, Evaluate, Analyse and Create when compared to Understand and Apply.

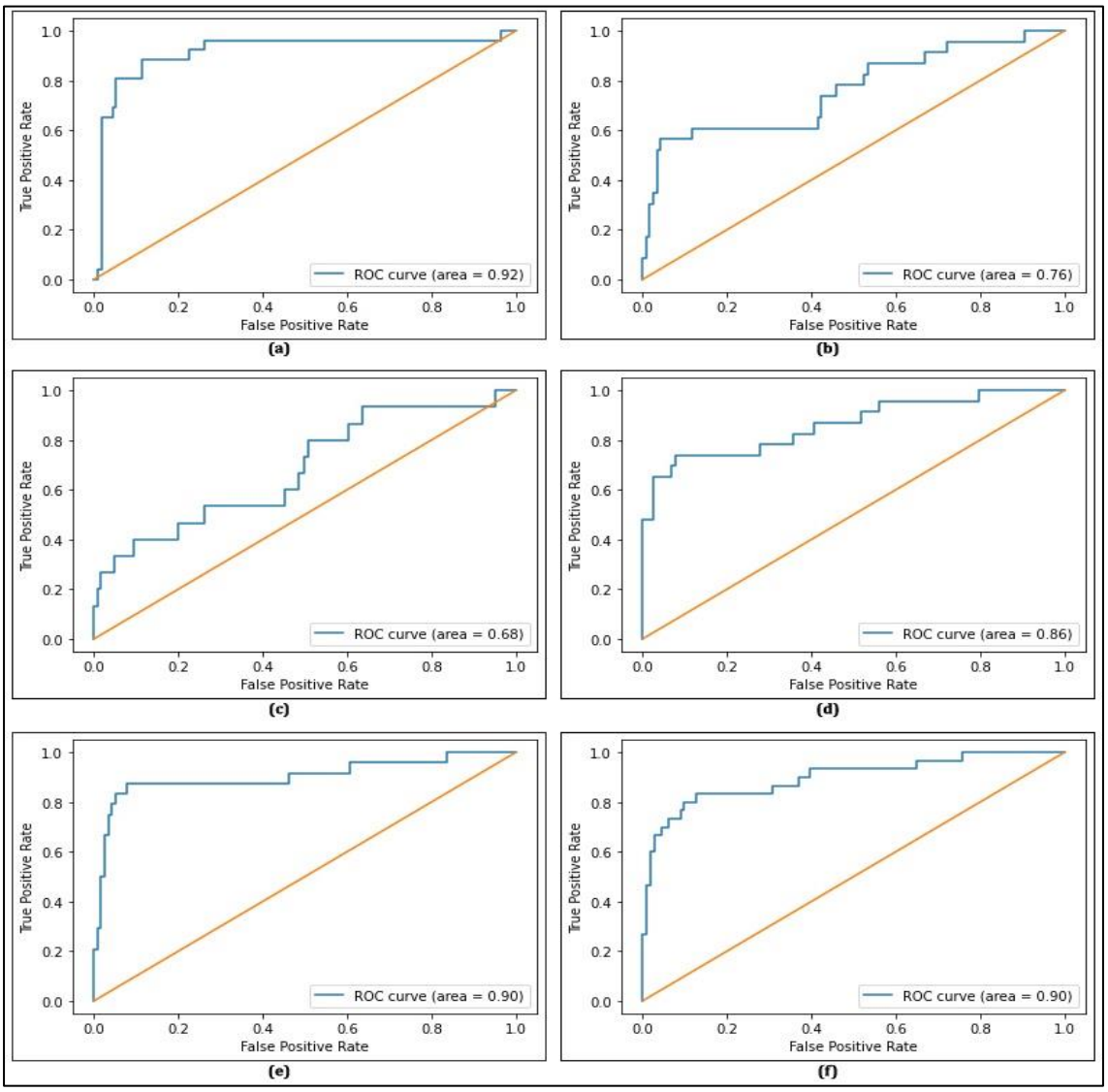

Figure 5: ROC curve for (a) Remember, (b) Understand, (c) Apply, (d) Analyse, (e) Evaluate and (f) Create for LSTM+BERT model

Performance Comparison of ANN+ TF-IDF and LSTM with GloVe and BERT

To identify the best approach for classifying the levels of revised Bloom's taxonomy, three models were developed and compared against each other: ANN with TF-IDF, LSTM with GloVe, and LSTM with BERT. Evaluation metrics such as accuracy, loss, MSE, MAE, precision, recall, and F1-score were used to compare the models' performance.

Figure 6 shows the variation in accuracies obtained for the selected models. The accuracy level increases with the increment in the number of epochs. Despite this, the highest classification accuracy was achieved for LSTM+BERT.

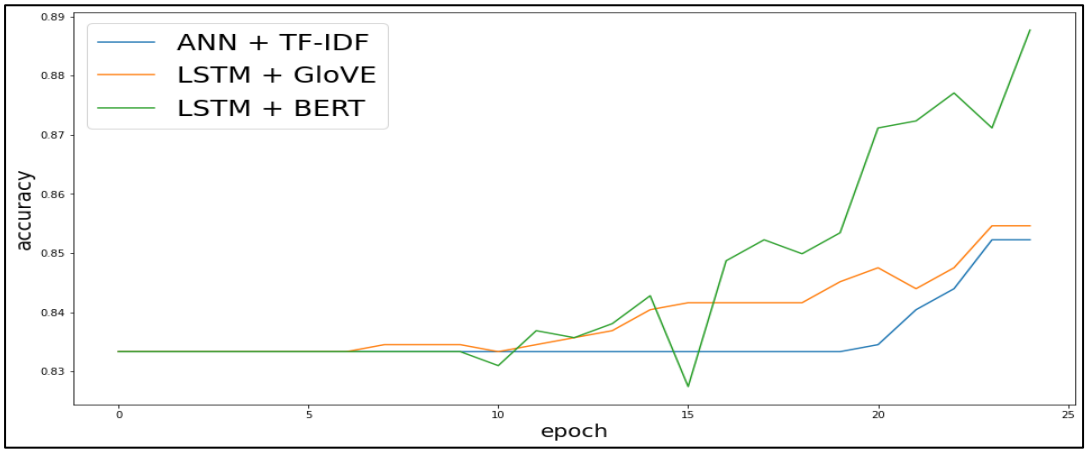

Figure 6: Variation in accuracies obtained for the models

In addition to evaluating the variation in accuracies, the loss values were also compared among the models. Figure 7 shows the comparison graph of the loss values from the models. The loss function graphs follow the opposite pattern of the accuracy graphs, with the loss values decreasing for all models as the number of epochs increases. The LSTM+BERT approach achieved the minimum loss values.

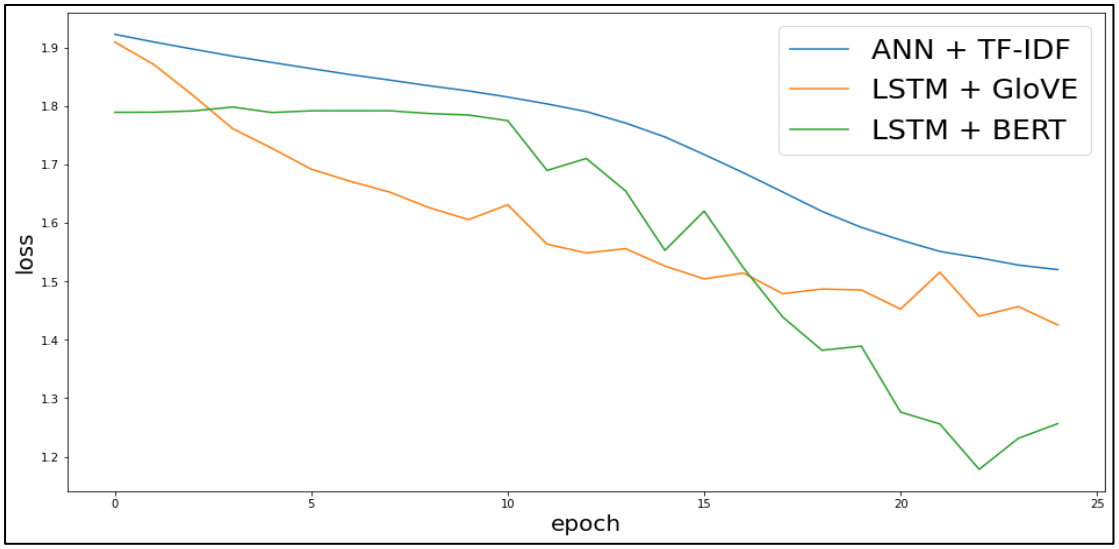

Figure 7: Variation in loss values obtained for the models

The MSE and MAE values were also obtained and used to identify the best approach. The graphs for the MSE and MAE values against the epochs are shown in Figures 8 and 9, respectively. As the number of epochs increases, both MAE and MSE values decrease. The LSTM+BERT approach achieved the minimum values for both MSE and MAE.

When the MSE is less, it suggests that the model is more accurate in its prediction of the variable being studied. Nevertheless, MSE is very sensitive to outliers and may be affected by values in the sample that are at the extreme end of the scale. Compared to MSE, MAE is a more reliable indicator of model performance since it is less susceptible to extreme values in the data. A more accurate prediction of the target variable may be inferred from a model by examining the MAE and looking for it to be lower. Both the mean square error (MSE) and mean absolute error (MAE), provide valuable insights into the accuracy of a model's predictions.

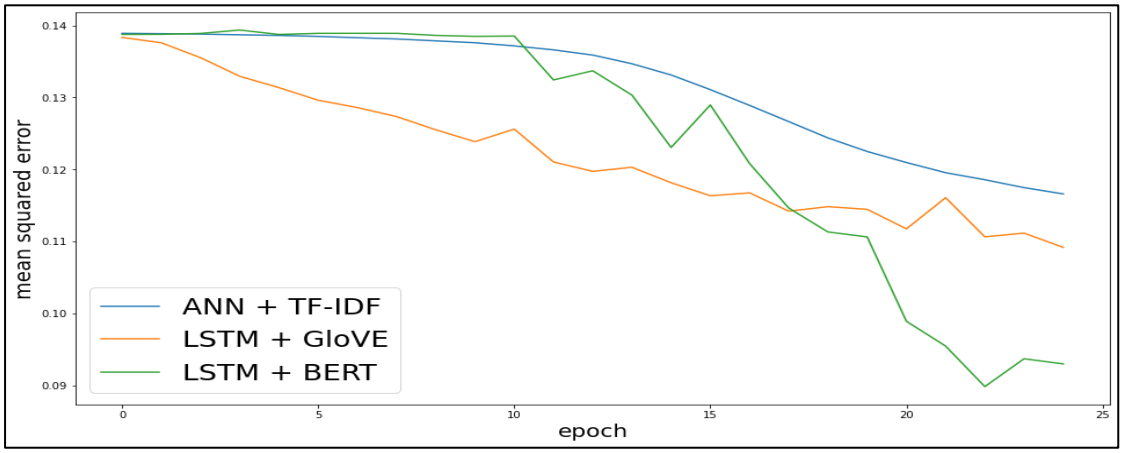

Figure 8: MSE for the implemented models

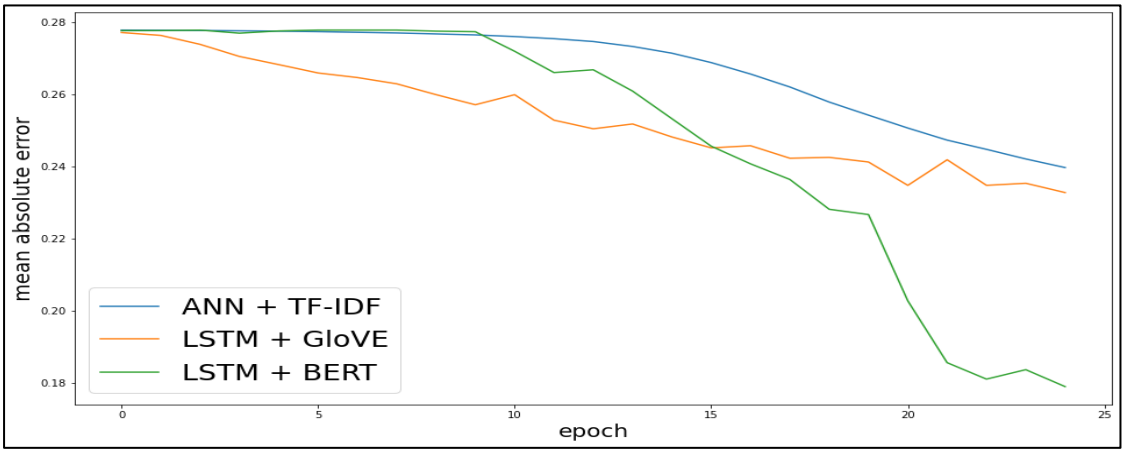

Figure 9: MAE for the implemented models

In addition to the abovementioned findings, Table 6 represents the other evaluation results obtained for the three models. Based on the results, the LSTM +BERT approach produced the results with the highest classification accuracy, precision, recall, and F1-values compared to the other three methods considered.

Table 6: *Comparison Results of Three Models*

Model	Accuracy	Precision	Recall	F1-Score
ANN + TF-IDF	85.54	80.36	15.98	25.24
LSTM + GloVe	85.78	91.02	14.44	24.03
LSTM + BERT	88.70	79.44	34.50	48.75

To compare the means of the accuracy scores of all three models and determine any significant differences, an analysis of variance (ANOVA) (Rao, 1992) was conducted, followed by post hoc pairwise comparisons using Bonferroni correction (Holm and Christman, 1985). ANOVA will tell whether there is a significant difference in accuracy scores among the three models, and post hoc comparisons will tell which models are significantly different from each other after controlling for multiple comparisons.

ANOVA results, F(2, 3) = 1179.12, p 0.05, indicate that there is a statistically significant difference in the performance of the three models (ANN + TF-IDF, LSTM + GloVe, and LSTM + BERT) on the supplied dataset. This indicates that at least one model has a substantially different mean performance score than the rest. Post hoc pairwise comparisons using the Bonferroni correction revealed a statistically significant difference between the three models' mean accuracy scores.

- The mean accuracy of LSTM + GloVe (M = 85.78, SD = 1.04) was not significantly different from that of ANN + TF-IDF (M = 85.54, SD = 1.03), $p > 0.05$.
- The mean accuracy of LSTM + BERT (M = 88.70, SD = 0.82) was significantly higher than that of ANN + TF-IDF (M = 85.54, SD = 1.03), $p < 0.001$.
- The mean accuracy of LSTM + BERT (M = 88.70, SD = 0.82) was significantly higher than that of LSTM + GloVe (M = 85.78, SD = 1.04), $p < 0.001$.

The Bonferroni correction for post hoc pairwise comparisons revealed that the mean accuracy of LSTM + BERT was considerably more significant than that of ANN + TF-IDF and LSTM + GloVe. Nevertheless, there was no significant difference between ANN + TF-IDF and LSTM + GloVe in terms of mean accuracy.

In conclusion, the ANOVA and post hoc pairwise comparisons indicate that the LSTM + BERT model performed considerably better in terms of accuracy than both the ANN + TF-IDF and LSTM + GloVe models. Similarly, there was no statistically significant difference between the ANN + TF-IDF and LSTM + GloVe models.

Significant disparities in the performance of various word embedding approaches and machine learning models show that the choice of word embedding technique and machine learning model may substantially affect the precision and dependability of the classification process. Specifically, the BERT embedding and LSTM models outperformed the other methods and models. This shows that deep learning and different word embedding approaches may improve the precision and dependability of the classification process.

Overall, our suggested approach has the potential to vastly enhance the efficacy and accuracy of educational assessment processes and provide valuable insights into the learning outcomes of students. This findings of this research can be used for implementation of the model in various educational contexts, including e-learning platforms and standardised testing. It can provide more personalised and objective evaluations of students' learning outcomes and improve the quality of education.

CONCLUSIONS AND RECOMMENDATIONS

In this study, a machine learning-based approach was proposed to classifying exam questions based on the revised Bloom's taxonomy levels. The proposed approach employed word embedding techniques such as BERT, TF-IDF, and GloVe, in combination with LSTM and ANN models, to classify the questions. A total of 16,584 questions were analysed and categorised into six levels of

the revised Bloom's taxonomy. The questions were pre-processed by removing stop words, symbols, quotes, tokenisation, and stemming. With the lowest loss, MAE, and MSE values, the LSTM+BERT model obtained the maximum classification accuracy of 88.7%. LSTM+BERT outperformed the other models, ANN+TF-IDF and LSTM+GloVe, according to the evaluation findings.

This research offers a viable classification method for exam questions based on the revised levels of Bloom's taxonomy. However, there is still an opportunity for advancement. To evaluate its potential to generalise, the suggested method may be used to categorise issues in many disciplines, such as science, mathematics, and literature. By examining the revised Bloom's taxonomy levels of questions presented to students before and after the teaching process, the suggested technique may also be used to evaluate the success of teaching approaches. In addition, the suggested method may be used to create an automated system that creates exam questions at a certain level of the revised Bloom's taxonomy. Educators may use this technology to generate exams with questions covering all levels of Bloom's taxonomy.

ACKNOWLEDGEMENTS

The authors would like to express sincere gratitude and appreciation to the team of 8 members who contributed to this research. This team, consisting of domain experts and authors, played a crucial role in developing our methodology by individually reviewing each question to assess its appropriateness concerning Bloom's taxonomy categories. Their dedication and expertise ensured the accuracy and reliability of our results, and we are deeply grateful for the contributions. We also extend our thanks to all those who supported and encouraged us throughout this research journey.

REFERENCES

Abduljabbar, D. A. & Omar, N. (2015). Exam Questions Classification Based on Bloom's Taxonomy Cognitive Level Using Classifiers Combination. *Journal of Theoretical and Applied Information Technology,* vol.78**,** 447.

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R. & Pintrich, P. (2001). A revision of Bloom's taxonomy of educational objectives. *A Taxonomy for Learning, Teaching and Assessing. Longman, New York*.

Aninditya, A., Hasibuan, M. A. & Sutoyo, E. (2019). Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy. 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), 2019. IEEE, pp.112-117.

Bloom, B. S. (1956). Taxonomy of Educational Objectives. Vol. 1: Cognitive Domain. *New York: McKay,* vol. 20, no. 1.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Forehand, M. (2005). Bloom's taxonomy: Original and revised. *Emerging perspectives on learning, teaching, and technology,* vol. 8**,** pp. 41-44.

Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.

Holm, K. & Christman, N. J. (1985). Post hoc tests following analysis of variance. *Research in nursing & health,* vol. 8**,** pp. 207-210.

Huang, J., Zhang, Z., Qiu J., Peng, L., Liu, D., Han, P. & Luo, K. (2021). Automatic Classroom Question Classification Based on Bloom's Taxonomy. 2021 13th International Conference on Education Technology and Computers, 2021. pp. 33-39.

Ifham, M., Banujan, K., Kumara, B. S. & Wijeratne, P. (2022). Automatic Classification of Questions based on Bloom's Taxonomy using Artificial Neural Network. 2022 International Conference on Decision Aid Sciences and Applications (DASA), 2022. IEEE, pp. 311-315.

Jain, M., Beniwal, R., Ghosh, A., Grover, T. & Tyagi U. (2019). Classifying Question Papers With Bloom's Taxonomy Using Machine Learning Techniques. Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3, 2019. Springer, pp. 399-408.

Jayakodi, K., Bandara, M. & Meedeniya, D. (2016). An automatic classifier for exam questions with WordNet and Cosine similarity. 2016 Moratuwa engineering research conference (MERCon), 2016a. IEEE, pp. 12-17.

Jayakodi, K., Bandara, M., Perera, I. & Meedeniya, D. (2016b). Wordnet and Cosine Similarity Based Classifier of Exam Questions Using Bloom's Taxonomy. *International Journal of Emerging Technologies in Learning (Online),* vol. 11**,** p. 142.

Kusuma, S. F., Siahaan, D. & Yuhana, U. L. (2015). Automatic Indonesia's questions classification based on bloom's taxonomy using Natural Language Processing a preliminary study. 2015 International Conference on Information Technology Systems and Innovation (ICITSI), 2015. IEEE, pp. 1-6.

Mcculloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics,* vol. 5**,** pp. 115-133.

Mohammed, M. & Omar, N. (2018). Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF. *International Journal on Advanced Science, Engineering and Information Technology,* vol. 8**,** pp. 1679-1685.

Mohammed, M. & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS one,* 15**,** e0230442.

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A. & Zulkifli, R. (2012). Automated Analysis of Exam Questions According to Bloom's Taxonomy. *Procedia-Social and Behavioral Sciences,* vol. 59**,** pp. 297-303.

Osman, A. & Yahya, A. A. (2016). Classifications of exam questions using natural language syntatic features: A case study based on Bloom's taxonomy. Proc. 3rd Int. Arab Conf. Qual. Assurance Higher Educ, 2016. pp. 1-8.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. pp. 1532-1543.

Rao, C. R. (1992). RA Fisher: The founder of modern statistics. *Statistical science***,** pp. 34-48.

Santos, M. J., Medina, A., Mateos Roco, J. M. & Queiruga-Dios, A. (2021). Compartmental Learning versus Joint Learning in Engineering Education. *Mathematics,* vol. 9**,** p. 662.

Schmidhuber, J. & Hochreiter, S. (1997). Long short-term memory. *Neural Comput,* vol. 9**,** pp. 1735-1780.

Shaikh, S., Daudpotta, S. M. & Imran, A. S. (2021). Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings. *IEEE Access,* vol. 9**,** 117887-117909.

Sparck Jones, S, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation,* vol. 28**,** pp. 11-21.

Supriyanto, C., Yusof, N. & Nurhadiono, B. (2013). Two-level feature selection for naive bayes with kernel density estimation in question classification based on Bloom's cognitive levels. 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), 2013. IEEE, pp. 237-241.

Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.-L. & Hao, H.(2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing,* vol. 174**,** pp. 806-814.

Yahya, A. A. & Osman, A.(2011). Automatic Classification of Questions Into Bloom's Cognitive Levels Using Support Vector Machines. The International Arab Conference on Information Technology, Naif Arab University for Security Science (NAUSS), 2011. pp. 1-6.

Yahya, A. A., Osman, A., Taleb, A. & Alattab, A. A. (2013). Analysing the cognitive level of classroom questions using machine learning techniques. *Procedia-Social and Behavioral Sciences,* vol. 97**,** pp. 587-595.

Yahya, A. A., Toukal, Z. & Osman, A. (2012). Bloom's Taxonomy–based Classification for Item Bank Questions Using Support Vector Machines. *Modern advances in intelligent systems and tools.* Springer.

Yang, L., Li, C., Ding, Q. & Li, L. (2013). Combining lexical and semantic features for short text classification. *Procedia Computer Science,* vol. 22**,** pp. 78-86.

Zhang, J., Wong, C., Giacaman, N. & Luxton-Reilly, A. (2021). Automated Classification of Computing Education Questions using Bloom's Taxonomy. Australasian Computing Education Conference, 2021. pp. 58-65.