Intelligent Classification of Educational Questions Using XLNet-CNN and Deep Contextual Embeddings.

Abstract

The automated classification of examination questions according to Bloom's Taxonomy (BT) assists question setters in developing high-quality assessments by accurately categorising questions into cognitive levels. While most previous studies in this area have employed traditional machine learning methods, relatively few have explored deep learning-based approaches. Contextual embeddings, in particular, have shown effectiveness across various natural language processing tasks. This study aims to evaluate a hybrid optimal pre-trained contextual word embedding technique, XLNet,combined with a Convolutional Neural Network (CNN) model tailored for BT-based question classification. To this end, the study examines the performance of the proposed XLNet+ CNN model with state-of-the-art models. Experimental results indicate that the XLNet + CNN model offers performancecomparable to existing models, with the added advantage of higher precision athigher cognitive levels.

Keywords: XLNet, CNN, NLP, Bloom, Education technology, Deep learning, word embedding

1. INTRODUCTION

The classification of examination questions by cognitive level is an important part of educational assessment, which ensures that students' learning outcomes are effectively measured. Bloom's Taxonomy is a widely accepted framework in educational theory that provides a structured approach to categorising learning objectives into three primary domains, viz., *cognitive*, *affective*, and *psychomotor*. Among these, the cognitive domain deals with different levels of thinking skills, from basic knowledge recall to higher-order analysis, synthesis, and evaluation. These hierarchical levels provide guidelines for educators to design balanced assessments that align with students' cognitive abilities [1].

Instructors often rely on Bloom's Taxonomy to formulate exam questions. However, manually constructing questions that accurately reflect these levels can be time-consuming and subjective and might also lead to inconsistencies in assessing students' skills. Therefore, there has been increased interest in automating the classification of exam questions based on Bloom's Taxonomy [2].

Despite the potential benefits of automation, most studies have focused on traditional machine learning techniques and only a few have explored deep learning approaches [3]. Deep learning models based on pre-trained word embeddings have been proven successful in various natural language processing tasks, such as automating question classification [4].

This research work aims to address the challenge of automating the classification of exam questions by developing a deep learning model that leverages the pre-trained contextualised word embeddings XLNet, in conjunction with Convolutional Neural Networks (CNNs). CNNs are known for their ability to extract key features from text, and are well-suited for this task as they focus on identifying patterns within the question text that correspond to specific cognitive levels of Bloom's Taxonomy.

The paper is structured as follows: Section 2 presents the study's motivation and contribution. Related work is presented in Section 3. The methodology is discussed in Section 4. Section 5 discusses the results, and Section provides the conclusion and future scope

2 MOTIVATION AND CONTRIBUTION

In today's fast-growing world, there is a requirement for high-quality educational assessments. Designing such assessments that align with Bloom's taxonomypresents significant challenges.Because it requires educators to carefully make the question paper that assessesvarious levels of student understanding according to the action verbs. Therefore, manually mapping the question papers to Bloom's taxonomy is difficult.

Given the requirements of higher education institutions and advancements in NLP and DL, there is a strong need for automated solutions to help teachers objectively classify examination questions [5]. This process of automation can save teachers' effort, improve consistency,and help to alignlearning outcomes and assessment strategies. The current work proposes a novel automated system for classifying examination questions based on Bloom's Taxonomy using advanced DL techniques. The key contributions are as follows:

1. To classify examination questions based on Bloom's Taxonomy, with the help of a novel hybrid deep learning approach that integrates contextual embeddings from XLNet with CNN.

2. The paper also compares the proposed model with the existing state-of-the-art models.

3. The proposed framework has been evaluated on the dataset and shows comparable performance.

3. RELATED WORK

With advancements in Natural Language Processing and Deep Learning, text classification tasks have gained significantprominence. Any text classification task relies on word embeddings, which transform textual data into numerical representations [6]. The literature suggests that earlier studies employed traditional Machine Learning (ML) approaches using feature selection techniques. However, with the emergence of deep learning models, neural network-based methods such asConvolutional Neural Networks (CNNs) have come to prominence in sentence classification tasks[7]. Furthermore, Yoon K (2014) highlighted that CNNs can automatically learn hierarchical features from text with minimal feature engineering [8]. CNNs are also

computationally efficient compared to sequential models such as LSTMs, which makes them suitable for large-scale applications [9].

In recent years, transformer-based models such as RoBERTa and XLNet have gained prominence. This is because they can generate contextualised word embeddings. These models are pre-trained on large corpora and can capture semantic and syntactic relationships more effectively than traditional embedding techniques like Word2Vec or GloVe [10].

Studies have shown that transformer-based embeddings significantly improve performance in various NLP tasks[11].In the context of education, many recent studies have explored the use of pre-trained models to classify examination questions according to Bloom's Taxonomy. These approaches emphasise the importance of selecting appropriate embedding techniques to capture contextual meaning and improve model effectiveness [12].

Despite these advancements,there are many research gaps. Many existing studies rely on either traditional embeddings or standalone deep learning models. The literature shows limited exploration of hybrid architectures that combine transformer-based embeddings with CNNs [13].

To address these limitations, recent research trends focus on integrating contextual embeddings from transformer models with efficient feature extraction mechanisms such as CNNs. Such hybrid approaches aim to leverage the strengths of both models, i.e., contextual understanding from transformers and spatial feature extraction from CNNs,to achieve higher accuracy and better generalisation.In this context, the present work builds upon prior studies by combining XLNet embeddings with CNN-based architectures to develop an efficient and accurate system for classifying examination questions according to Bloom's Taxonomy and comparing its performance with the state-of-the-art RoBERTa+CNN model as proposed by [12]

4. METHODOLOGY

This section presents the methodology for implementing the proposed framework.

4.1. Dataset

In this study, five datasets were utilised, comprising both previously established datasets and one collected specifically for this research. The dataset collected contained 1,200 questions, all of which were labelled by educators. After merging all five datasets, a comprehensive dataset of 2,522 questions was formed.Ultimately, the proposed models were trained and evaluated exclusively on the combined dataset as given in [12] . Figure 1 presents the percentage of the number of questions against each Bloom's category.Figure 2 presents word clouds of various classes containing action verbs.

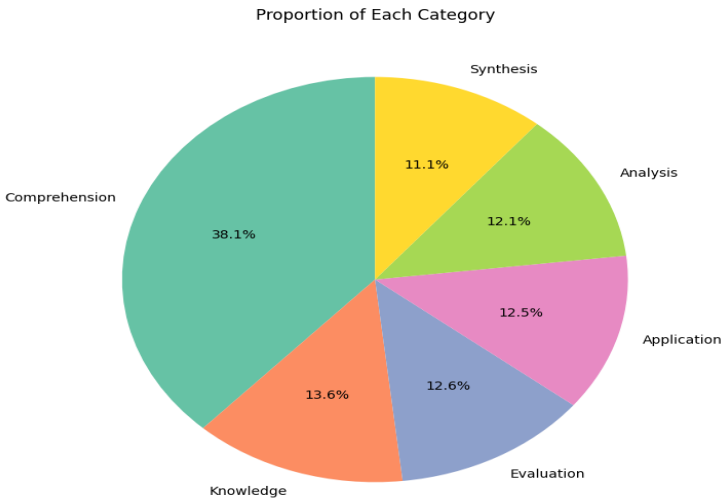
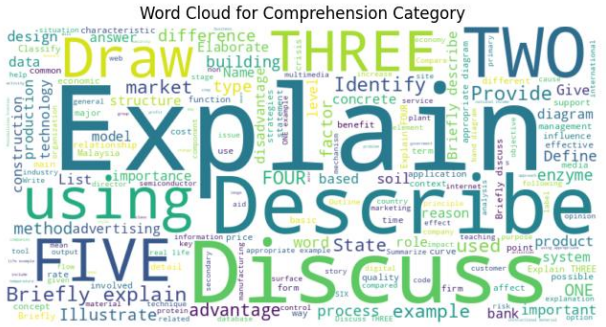
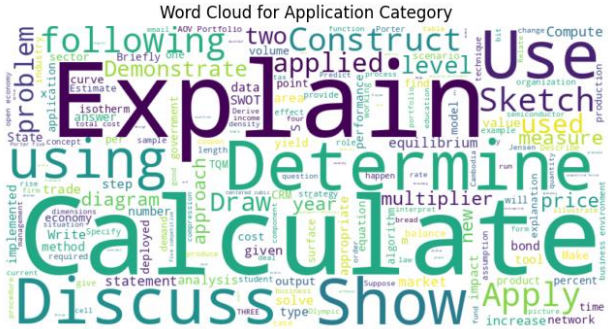
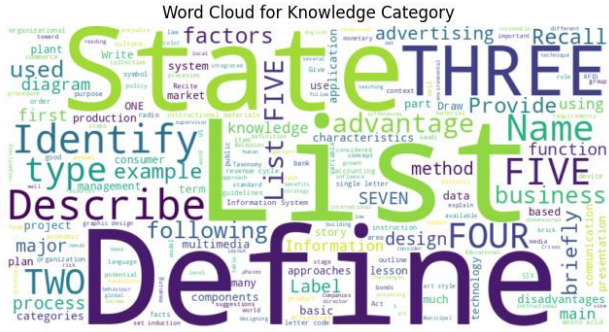


Figure 1. Percentage of Number of questions in each class in the dataset







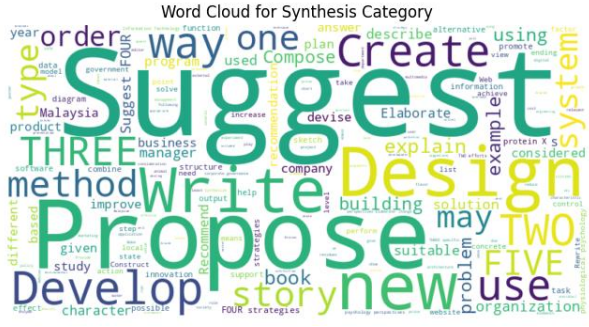


Figure 2 Word cloud of various classes

4.2Proposed Framework (XLNet+CNN)

Algorithm 1 below provides the key details of the proposed framework.

Input:
Dataset of examination questions

Output:
Predicted Bloom's Taxonomy category

Step 1: Data Loading and Preparation

1. *Load dataset from Excel file.*
2. *Assign numerical labels to Bloom's Taxonomy categories.*
3. *Preprocess text data (cleaning, normalization).*
4. *Tokenize text using XLNet tokenizer:*
    - *Convert text into input IDs*
    - *Generate attention masks*
    - *Set fixed sequence length (e.g., 256 tokens)*

Step 2: Tokenization and Embedding Generation

5. *Store tokenized outputs (input IDs and attention masks) in arrays.*
6. *Pass tokenized inputs to XLNet to generate contextual embeddings.*

Step 3: Model Architecture Design

7. *Input XLNet embeddings into Convolutional Neural Networks (CNN):*
    - *Apply Conv1D layers to extract features*
    - *Apply Max-Pooling layers to reduce dimensionality*
8. *Flatten the output of CNN layers.*
9. *Pass flattened output to Dense layer with ReLU activation.*
10. *Add output layer with Softmax activation to classify into six categories.*

Step 4: Training and Validation

11. *Split dataset into training and validation sets.*
12. *Compile model using:*

- *Optimizer: Adam*
- *Loss function: Categorical Cross-Entropy*

- *Metric: Accuracy*

13. *Train model for a fixed number of epochs (e.g., 5 epochs).*
14. *Store training and validation accuracy and loss.*

Step 5: Evaluation and Classification

15. *Evaluate model performance on validation dataset.*
16. *Visualize accuracy and loss trends.*
17. *Define classification function:*

- *Tokenize new question using XLNet tokenizer*
- *Pass through trained model*
- *Obtain Softmax probabilities*
- *Select highest probability class*

18. *Map predicted label to Bloom's Taxonomy level.*

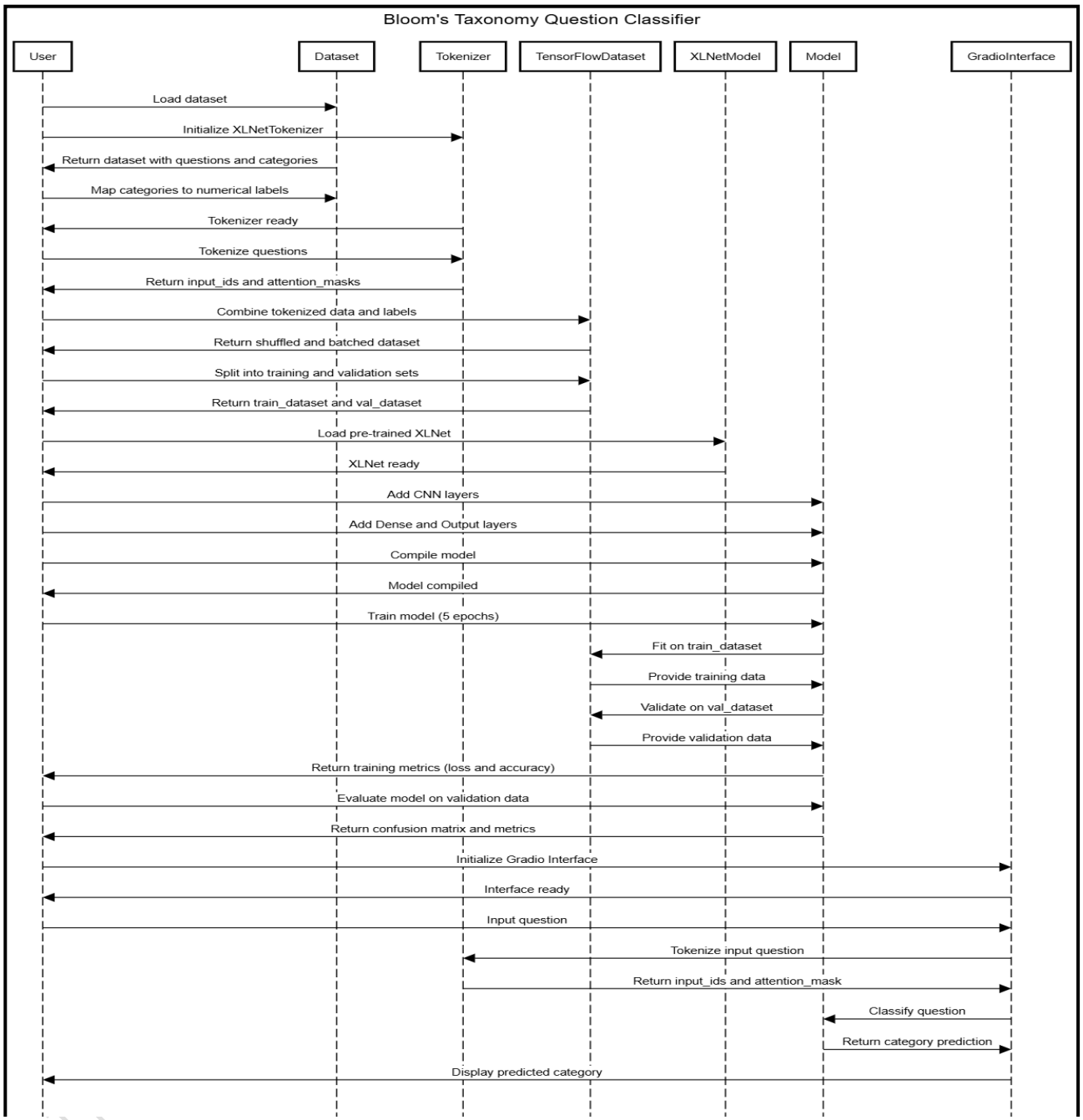


Figure 3: Sequence diagram of the proposed framework implementation

Hyperparameter tuning for the proposed framework have been done as represented in Table 2

Table 2: Hyperparameters for XLNetCNN:

Hyperparameter	**Value**
Text Preprocessing	Tokenizer: `XLNetTokenizer`
Max Sequence Length	256
Input IDs	Shape: `(batch size, 256)`

Attention Masks	Shape: `(batch_size, 256)`
Model Type	XLNet (`xlnet-base-cased`)
Embedding Layer	`TFXLNetModel`
Conv1D Layer Filters	128
Conv1D Kernel Size	3
Conv1D Activation	`ReLU`
MaxPooling1D Pool Size	2
Second Conv1D Filters	64
Second Conv1D Kernel Size	3
Second Conv1D Activation	`ReLU`
Fully Connected Layer	Dense (256 units, `ReLU`)
Output Layer	Dense (6 units, `Softmax`)
Optimizer	Adam (`learning_rate=1e-5`)
Loss Function	Categorical Crossentropy
Metrics	Categorical Accuracy
Batch Size	16

5. RESULT AND DISCUSSION

The results of this study highlight the significant impact of word embedding techniques on the classification of examination questions into Bloom's Taxonomy (BT) categories. **XLNet+CNN** model was trained and validated on a dataset of 2,522 questions, with performance evaluated using Accuracy and F1 Score metrics.The classification report is shown in Figure 4, and the confusion matrix in Figure 5.

```
Classification Report:
                precision    recall  f1-score   support

     Knowledge       0.92      0.80      0.86        60
Comprehension        0.84      0.95      0.89       212
   Application       0.79      0.63      0.70        60
      Analysis       0.94      0.78      0.85        58
     Synthesis       0.76      0.88      0.82        48
    Evaluation       0.94      0.84      0.89        58

      accuracy                           0.85       496
     macro avg       0.87      0.81      0.83       496
  weighted avg       0.86      0.85      0.85       496
```

Figure 4 Classification Report of the proposed Framework

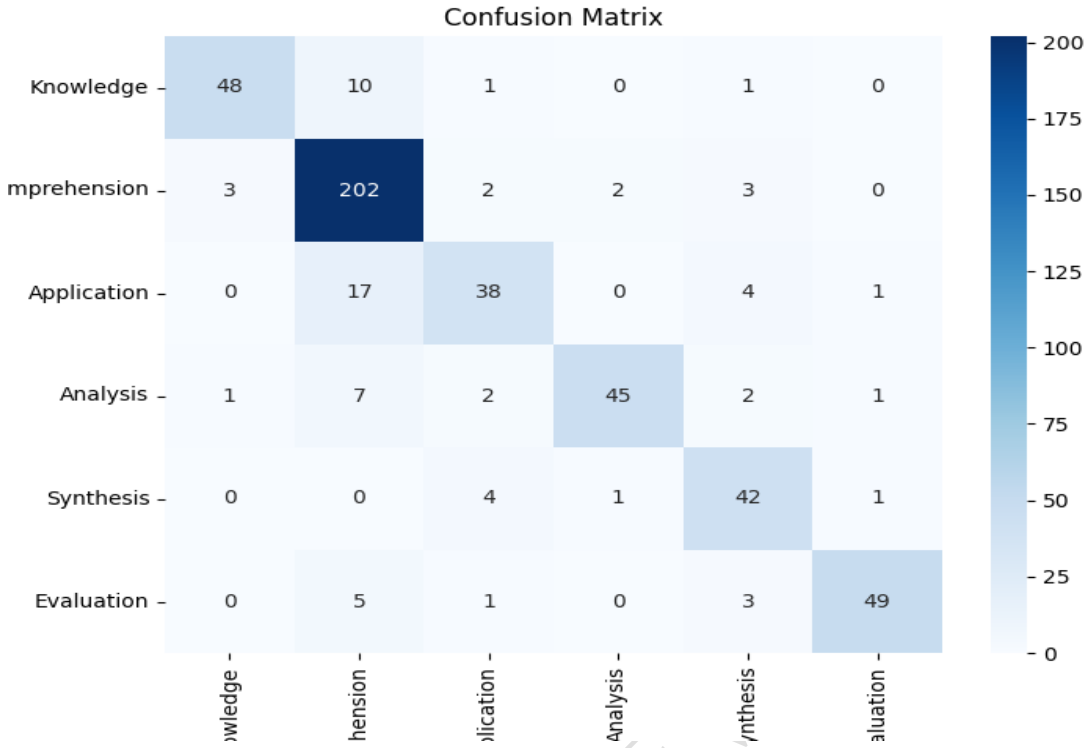


Figure 5: Confusion matrix of the proposed Framework

5.1. Comparative Performance Analysis

The performance of the proposed framework XLNET +CNN) has been compared with the state-of-the-art best RoBERTa +CNN as proposed by. Table 3 presents the results below, and Table 4 presents the class-wise performance of both models in terms of precision, recall, and F1 Score.

Table 3 Comparative Performance of the proposed framework

Technique	Accuracy	F1-score
RobertaCnn (Gani,M.)	86.2%	86.1%
xlne+CNN (Proposed Model)	85.63%	84.03%

Table 4:Cla-s wise performance

Class	XLNet + CNN (P / R / F1)	RoBERTa + CNN (P / R / F1)
Knowledge	0.92 / 0.80 / 0.86	0.91 / 0.86 / 0.88
Comprehension	0.84 / 0.95 / 0.89	0.89 / 0.93 / 0.91
Application	0.79 / 0.63 / 0.70	0.85 / 0.72 / 0.78
Analysis	0.94 / 0.78 / 0.85	0.90 / 0.87 / 0.88

Class	XLNet + CNN (P / R / F1)	RoBERTa + CNN (P / R / F1)
Synthesis	0.76 / 0.88 / 0.82	0.72 / 0.93 / 0.81
Evaluation	0.94 / 0.84 / 0.89	0.86 / 0.75 / 0.80

The comparative analysis of the XLNet + CNN and RoBERTa + CNN models indicates that both approaches exhibit comparable performance in classifying examination questions according to Bloom's Taxonomy. While the RoBERTa + CNN model shows a slightly higher overall accuracy and F1-score, the difference in performance is marginal, suggesting that both models are equally effective for this task.XLNet + CNN model demonstrates certain advantages at the class level. It achieves higher precision in higher-order cognitive categories such as Analysis and Evaluation. Additionally, XLNet shows competitive performance in handling the complex semantic structures of higher-order thinking-skills questions due to its permutation-based contextual learning mechanism.Overall, the findings suggest that RoBERTa + CNN performs marginally better in terms ofAccuracy and F1 Score. However, the XLNet + CNN model offers comparable performance while providing the added advantage of higher precision athigher cognitive levels**.**

6. CONCLUSION AND FUTURE WORK:

This study evaluated the performance of the proposed XLNet + CNN model for automatically classifying examination questions by Bloom's Taxonomy and compared it with the state-of-the-art RoBERTa + CNN model. The results indicated that both models achieved comparable performance. Integrating automated question-generation systems with educational paradigms advances intelligent assessment systems in education.

This study has certain limitations, including a relatively small dataset and class imbalance, which may affect the model's generalizability and bias the performance toward dominant categories. This approach hasn't considered the contextual metadata, as it was limited to textual features. The future work can focus on enhancing the robustness of the proposed approach. This can be done using larger and more diverse datasets. Also, incorporating explainable AI techniques will improve transparency and trust in the system.

REFERENCES

[1] Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, *103*(3), 152.

[2] Banujan, K., Kumara, S., Prasanth, S., & Ravikumar, N. (2023). Revolutionising Educational Assessment: Automated Question Classification Using Bloom's Taxonomy and Deep Learning Techniques--A Case Study on Undergraduate Examination Questions. *International Journal of Education and Development using Information and Communication Technology*, *19*(3), 259-278.

[3] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1-41.

[4] Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial intelligence review*, *56*(9), 10345-10425.

[5] Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., ... &Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in social and administrative pharmacy*, *19*(8), 1236-1242.

[6] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, *54*(3), 1-40.
[7] Soni, S., Chouhan, S. S., & Rathore, S. S. (2023). TextConvoNet: A convolutional neural network based architecture for text classification. *Applied Intelligence*, *53*(11), 14249-14268.
[8] Lee, Y., Yoon, S., & Jung, K. (2018, October). Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd workshop on abusive language online (ALW2)* (pp. 101-106).
[9] Shen, L., Sun, Y., Yu, Z., Ding, L., Tian, X., & Tao, D. (2024). On efficient training of large-scale deep learning models. *ACM Computing Surveys*, *57*(3), 1-36.
[10] Areshey, A., &Mathkour, H. (2024). Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. *Expert Systems*, *41*(11), e13701.
[11] Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58-69.
[12] Gani, M. O., Ayyasamy, R. K., Sangodiah, A., & Fui, Y. T. (2023). Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique. *Education and Information Technologies*, *28*(12), 15893-15914.
[13] Wu, H., Liu, Y., & Wang, J. (2020). Review of Text Classification Methods on Deep Learning. *Computers, Materials & Continua*, *63*(3).