ISSN (O): 2320-5407
ISSN (P): 3107-4928

Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/23135
DOI URL: http://dx.doi.org/10.21474/IJAR01/23135

RESEARCH ARTICLE

AN EMPIRICAL EVALUATION OF CONTEXTUAL EMBEDDING-BASED MODELS FOR CLASSIFICATION OF EDUCATIONAL QUESTIONS USING BLOOM'S TAXONOMY

Parneet Kaur

Assistant Professor, Central University of Punjab, Bathinda, India.

Manuscript Info

Manuscript History
Received: 16 January 2026
Final Accepted: 18 February 2026
Published: March 2026

Keywords:-
XLNet, CNN, NLP, Bloom, Education technology, Deep learning, WordEmbedding,Bloom's Taxonomy, Question Classification

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 achieves performance comparable to existing models. Although it is 0.5% lower in overall accuracy than RoBERTa + CNN, it has 8% higher precision for the higher-order cognitive skill Evaluation category and 4% higher precision for the Analysis Category. Despite slightly lower accuracy, XLNet + CNN demonstrates superior precision and better identification of higher-order cognitive skills, making it more suitable for reliable educational assessment tasks.

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

Corresponding Author:-Parneet Kaur
Address:-Central University of Punjab.
kaurparneet0410@gmail.com

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.

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 assesses various 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.

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 across 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]. The literature also reveals that, unlike RoBERTa, which relies on masked language modelling and may lose contextual information during training [14], XLNet employs a permutation-based autoregressive objective that captures bidirectional context without masking [15]. This enables more complete and context-rich representations, which, when combined with CNNs, improve feature extraction for complex classification tasks [16] .

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 study combines 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].

Methodology:-

This section presents the methodology for implementing the proposed framework.

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 described 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 for each Bloom's category.Class imbalance was addressed through loss function reweighting, assigning higher weights to minority classes to reduce bias toward majority classes and improve model learning.

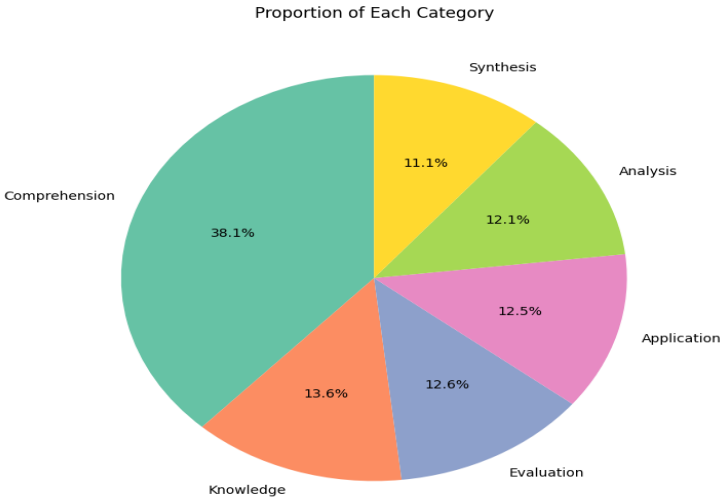

Figure 1. Percentage of the number of questions in each class in the dataset

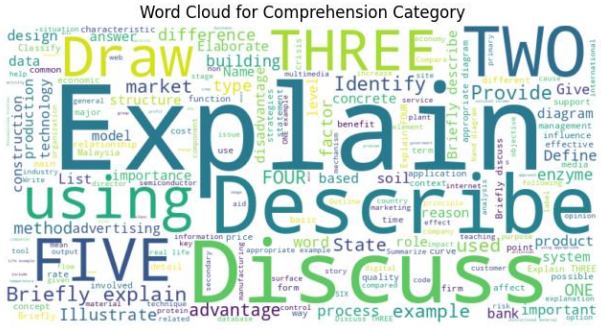

Figure 2 Word cloud of various classes

Proposed Framework (XLNet+CNN):-

Initially, the input dataset is preprocessed by performing text-cleaning steps such as noise removal, tokenisation, and normalisation. The cleaned text is then passed through the XLNet model, which generates deep contextual embeddings by capturing bidirectional dependencies using its permutation-based training mechanism. These embeddings serve as rich feature representations of the input text. After this, the extracted XLNet features are fed into a Convolutional Neural Network (CNN), which helps refine the feature space by capturing important n-gram-level information. The resulting feature maps are then passed through fully connected layers, and a softmax classifier is applied to predict the final class labels corresponding to Bloom's Taxonomy levels. The model is trained with an appropriate loss function. Class imbalance is addressed using weighted loss or similar techniques. Table 1 below provides the key details of the proposed framework**.**

Table 1 Algorithm for the proposed framework

Input:
Dataset of examination questions

Output:
Predicted Bloom's Taxonomy category

1) Data Loading and Preparation
1.1 Load dataset and assign numerical labels to Bloom's Taxonomy categories.
1.2 Tokenize text using XLNet tokenizer:

2) Tokenization and Embedding Generation
2.1 Store tokenized outputs (input IDs and attention masks) in arrays.
2.2. Pass tokenized inputs to XLNet to generate contextual embeddings.

3) Model Architecture Design
3.1 Input XLNet embeddings into Convolutional Neural Networks (CNN):
3.2 Flatten the output of CNN layers.
3.3 Pass flattened output to Dense layer with ReLU activation.
3.4 Add output layer with Softmax activation to classify into six categories.

4). Training and Validation
4.1 Split dataset into training and validation sets.
4.2 Compile model using Adam Optimizer: *Adam* and *Categorical Cross-Entropy*
4.3 Train model and store training and validation accuracy and loss.

5). Evaluation and Classification
5.1 Evaluate model performance on validation dataset.
5.2 Visualize accuracy and loss trends.
5.3 Map predicted label to Bloom's Taxonomy level.

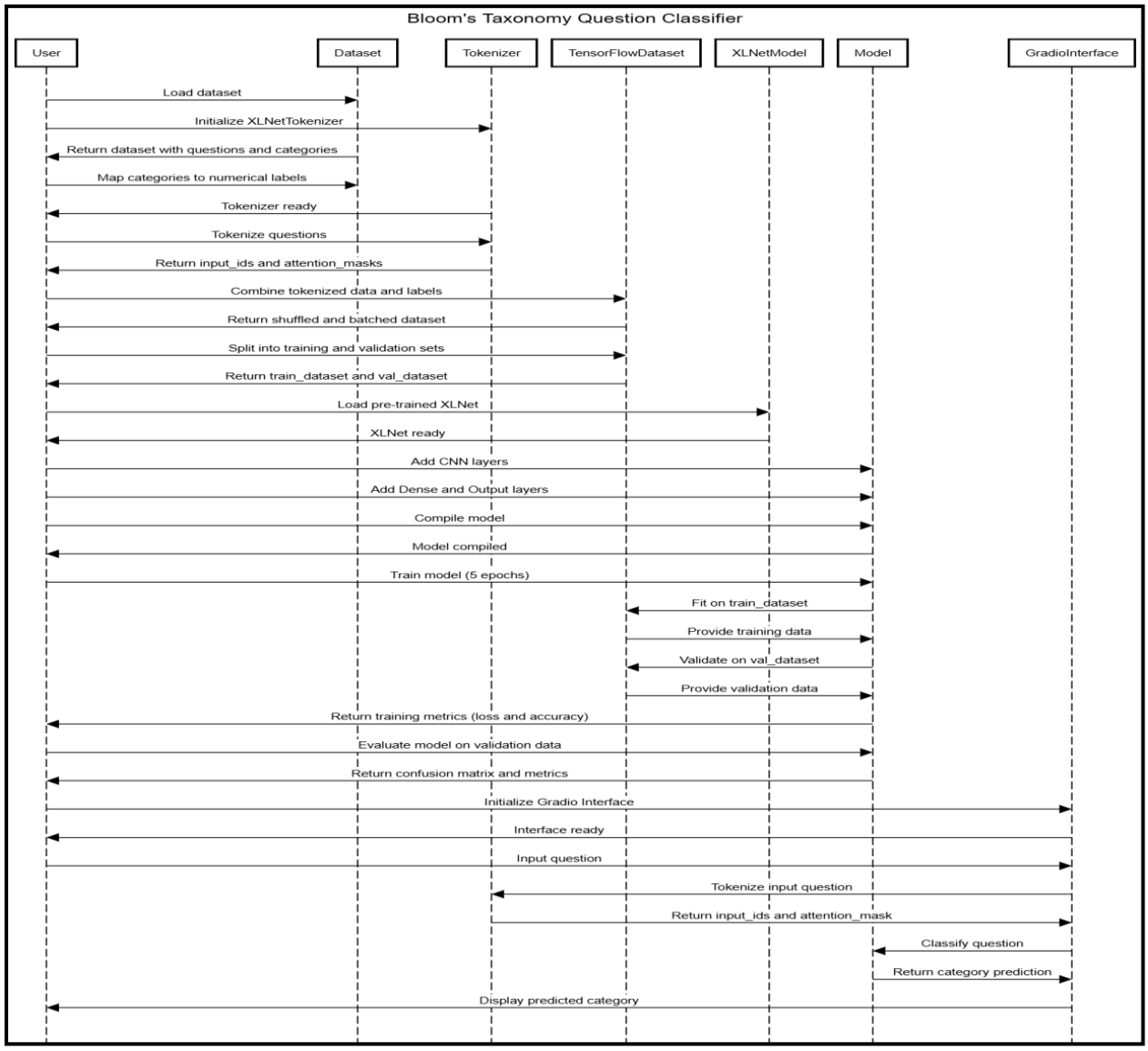

Figure 3: Sequence diagram of the proposed framework implementation

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

Table 2: Hyperparameters for XLNet+ CNN:

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
Pooling Strategy	MaxPooling1D
Pool Size	2
Stride	1
Padding	Valid
Second Conv1D Filters	64
Second Conv1D Kernel Size	3
Second Conv1D Activation	ReLU
Second Pooling Layer	MaxPooling1D (pool size = 2)
Flatten Layer	Applied
Fully Connected Layer	Dense (256 units, ReLU)
Output Layer	Dense (6 units, Softmax)
Dropout	0.4
Optimizer	Adam (learning_rate = 1e-5)
Loss Function	Categorical Crossentropy
Batch Size	16
Number of Epochs	5

Results 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. The 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

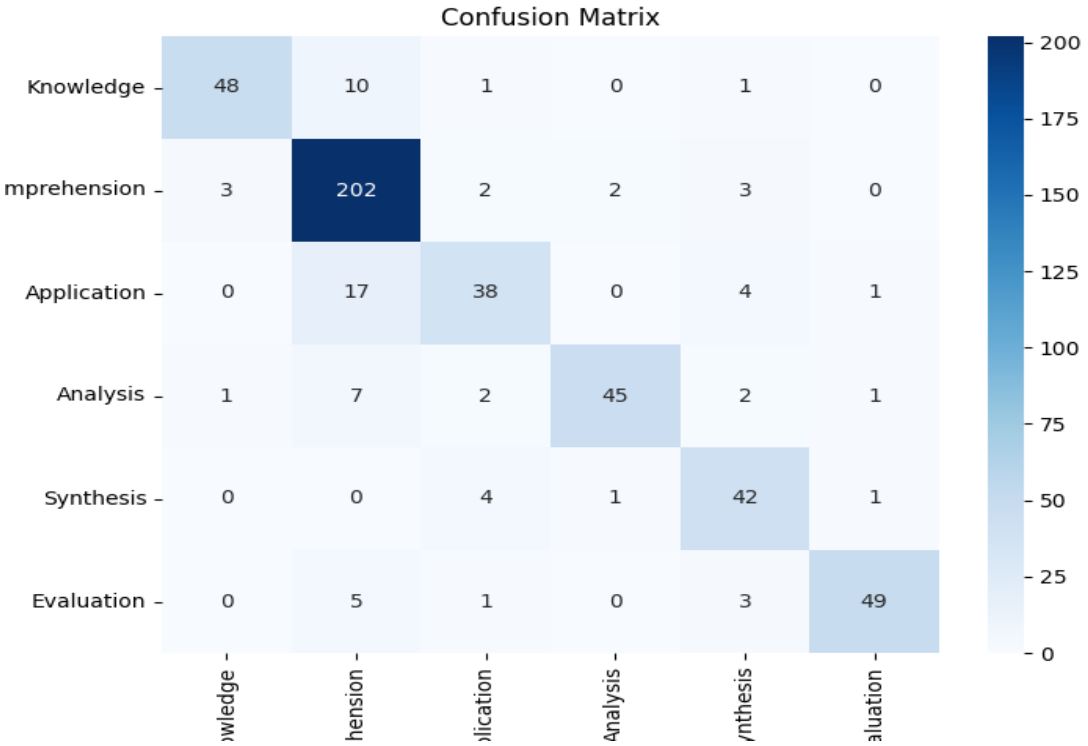

Figure 5: Confusion matrix of the proposed Framework

Comparative Performance Analysis:-

The performance of the proposed framework XLNET +CNN) has been compared with the state-of-the-art RoBERTa +CNN, as proposed by[12]. Table 3 presents the results in terms of accuracy, and F1-Score, and Table 4 presents the class-wise performance of both models in termsof precision, recall, and F1 Score.

Table 3: Comparative Performance of the Proposed Framework

Technique	Accuracy	F1-score
Roberta +CNN(Gani,M. et al., ,2023)	86.2%	86.1%
XLNet+CNN (Proposed Model)	85.63%	84.03%

Table 4:Class-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
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 experimental results indicate that the XLNet combined with the CNN model attains an overall accuracy of 85.63% and an F1-score of 84.03%, which are marginally lower than those of the RoBERTa+ CNN model, whichrecorded 86.2% accuracy and 86.1% F1-score. However, a more granular analysis by class reveals distinct advantages of the XLNet-based approach. Specifically, it achieves higher precision in critical categories such as Knowledge (0.92), Analysis (0.94), and Evaluation (0.94), indicating a lower rate of false-positive classifications in these groups. In addition, its F1-scores are competitive and surpass those of RoBERTa in certain classes, for

instance, Synthesis (0.82 compared to 0.81) and Evaluation (0.89 versus 0.80), underscoring its relative efficacy in processing more complex cognitive domains.

From an interpretive standpoint, these outcomes suggest that the XLNet and CNN architectures exhibit greater robustness in identifying higher-order cognitive skills, which are essential for tasks aligned with Bloom's Taxonomy. Its notably higher recall rates in Synthesis (0.88) and Evaluation (0.84) indicate an enhanced ability to capture a larger proportion of relevant instances in these demanding categories. While RoBERTa, combined with a CNN, shows marginally superior aggregate performance and excels at intermediate-level classifications, the XLNet-based model offers more consistent and dependable performance when addressing advanced cognitive levels. Consequently, this characteristic makes the XLNet+ CNN approach particularly advantageous in educational contexts where precise recognition of higher-order thinking skills is moreimportant than slight gains in overall performance metrics.

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 study is limited to a comparison between the XLNet+CNN hybrid architecture and the RoBERTa+CNN model. Other potential architectures, such as DistilBERT and similar lightweight transformer models, were not explored in this work. Incorporating and evaluating these alternative models could provide additional insights and may be considered as part of future research directions.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 can improve transparency and trust in the system.

Data availability: Gani, Mohammed Osman; Sangodiah, Anbuselvan (2023). Exam Question Datasets. figshare. Dataset. https://doi.org/10.6084/m9.figshare.22597957.v3

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).
[14] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019).
XLNet: Generalized autoregressive pretraining for language understanding.
Advances in Neural Information Processing Systems, 32. https://doi.org/10.48550/arXiv.1906.08237
[15] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019).RoBERTa: A robustly optimized BERT pretraining approach.
arXiv preprint. https://doi.org/10.48550/arXiv.1907.11692
[16] Kim, Y. (2014).Convolutional neural networks for sentence classification.
Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751. https://doi.org/10.3115/v1/D14-1181