International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Aug-2024)

http://dx.doi.org/10.12785/ijcds/160170

A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data

Elia Ahidi Elisante Lukwaro[1,2], Khamisi Kalegele[2] and Devotha G. Nyambo[1]

[1]*Department of ICSE,Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania*
[2]*Department of Mathematics and ICT, The Open University of Tanzania, Dar Es Salaam, Tanzania*

Received 12 May 2023, Revised 05 May 2024, Accepted 13 May 2024, Published 10 Aug. 2024

Abstract: There has been a significant increase in academic processes to ensure the quality of educational resources such as curricula, examinations, and educational content. This has drawn attention to studies exploring the use of text mining, learning machines, and auto-analytic tools like natural language processing (NLP) to interpret and evaluate the quality of these educational resources. Auto-analytical techniques are required to evaluate the quality of educational content; otherwise, manual evaluation can be burdensome and improperly influenced by human instincts. This study employs a methodical approach to comprehensively survey NLP techniques for extracting syntactic and semantic features to analyze and comprehend educational content. NLP, in combination with machine learning, is an ideal tool for automatically evaluating the aspects of higher education quality. This is because they include features that aid in textual content comprehension as well as implementing natural language techniques that provide an interpretive interface between humans and machines. The review highlights the limitations of NLP in evaluating educational data, including the need for sentence-level understanding and the need for research to address challenges like noise in text data, domain-specific language variations, and improving model robustness for effective feature extraction in educational contexts. The findings of this review hold substantial benefits for various stakeholders, including education regulatory bodies, researchers, higher education institutions, and NLP researchers. Notably, the study equips NLP researchers with valuable insights into document analysis's current strengths and weaknesses. The accumulated evidence can provide the skills to develop NLP-based applications for evaluating the relevant and quality aspects of education in higher educational settings. Furthermore, NLP researchers can be updated on the strengths and limitations of document analysis, allowing them to apply effective text representation approaches and implement the appropriate algorithm and techniques for NLP tasks, particularly in educational data.

Keywords: NLP, syntactic features, semantic feature, question classification, curriculum, educational content.

1. INTRODUCTION

Recently, statistics have shown a significant increase in higher education institute enrollment, while graduate unemployment is on the rise at both universities [1], [2], [3]. This trend has prompted a growing interest among researchers in leveraging technological advancements, such as natural language processing (NLP), to develop digital methods for evaluating educational processes linked to academic quality.

Higher education institutions (HEIs) are responsible for various academic processes that ensure their competitive survival and shape the quality of education. These processes produce structured and unstructured academic content, encompassing course materials, examination questions, programme specifics, and more. In this context, manual assessment of quality aspects is exceptionally challenging. Evaluating education quality is crucial for understanding the education system's effectiveness in fostering students' cognitive development, instilling values and attitudes, addressing local and global challenges, nurturing creativity and emotional growth, and fostering analytical problem-solving skills [4][5]. The quality of education is extremely complex because it involves many stakeholders with varying perspectives: the government, employers, academics, students, parents, and society at large, all of whom describe excellence differently [6]. However, the majority of universities globally present quality criteria or standards of education, such as educational activities, and the analysis of this can be evaluated through the following aspects: programmes/curriculum, assessments, admission system, and other resources [7], of which curriculum and assessment, i.e., examinations, are taken as a focal point for this study.

The mechanism to evaluate the aspects that determine the quality of education necessitates analytical, automated methods. Manual processes such as examination moderation, credit transfer, syllabus approval, and compliance can be burdensome and prone to errors if solely relied

E-mail address: elia.ahidi@out.ac.tz, khamisi.kalegele@out.ac.tz, devotha.nyambo@nm-aist.ac.tz.

on human intuition [8]. Approaches used to evaluate the quality aspects of education can be categorised as human-based and automated. Total Quality Management (TQM) is one of the human-based approaches for managing quality used in business and education. It entails a set of principles and norms for improving the services and products offered to customers [9]. However, global concerns about education quality and resource constraints have pushed higher education institutions to look for automated options such as NLP, text mining, and machine learning techniques. NLP, a subfield of artificial intelligence, strives to equip computers to comprehend text and spoken language, much like human beings. By imbuing computers with these capabilities, we can automate numerous quality assessment operations. Already, there are ongoing efforts to utilise techniques like text mining and data mining to verify academic content similarities, audit educational information, assess examination question standards, review syllabi, evaluate factors influencing student performance, visualise learning activities, and more [6], [10], [11], [12], [13]. With the digitization of relevant data and tools, applying NLP techniques becomes increasingly important for assessing, controlling, and evaluating. This review investigates NLP techniques and the associated challenges in extracting features to analyze and evaluate education quality, specifically syllabi and examinations. The paper's objective is to examine the current state of NLP applications in educational text analysis and delve into the strengths and limitations of these techniques. The paper specifically aims to address the following research questions:

- RQ1: What are commonly used NLP techniques for feature extraction in education data, especially in syllabi and examinations, to assess quality?
- RQ2: What are the strengths and limitations of these feature extraction techniques in educational quality assessment?
- RQ3: How have existing NLP metrics been adapted to suit the distinctive characteristics of educational content?

The primary objective of this research is to explore the application of NLP techniques in extracting features from educational data and addressing associated challenges. Feature extraction, often termed feature engineering, entails deriving meaningful information from natural language sources like text and audio. The resulting word representations or embeddings serve as inputs for machine learning models, enabling them to undertake specific tasks by comprehending the nuances of natural language. Given the inherent complexity of natural languages, feature extraction necessitates the utilization of diverse approaches, each characterized by distinct advantages and challenges

The remainder of the paper is organized as follows: Section 2 presents the literature review, Section 3 outlines the proposed approach, Section 4 discusses the results and provides a comprehensive analysis, and Section 5 draws the conclusion.

2. LITERATURE IN REVIEW

The use of automated analytical technology in educational data has garnered significant attention. Numerous reviews have delved into different aspects of education, providing valuable insights into applying natural language processing techniques for assessing educational data. These reviews have taken various approaches, adding empirical evidence to the existing body of knowledge.

The work of [14], which introduced an integrated approach to feature extraction encompassing keyword, headword, syntactic, and semantic extraction, exemplifies comprehensive review approaches. They applied this approach to classify questions containing keywords assigned to multiple levels of Bloom's taxonomy (BT). Anbuselvan Sangodiah et al. [14] studied statistical methods for question-answering systems, information finding, and educational environments. They used machine learning tools such as support vector machines (SVM) and other classifiers. The study recognised that semantic and syntactic extraction are important for getting accurate results with SVM classifiers in information retrieval and question-answering systems, but it noted the relatively lower performance in educational settings. The results of [15] show that using NLP features like lexical and semantic matching along with machine learning methods like SVM makes question-answering systems better at classifying things. However, the study highlighted the substantial impact of the dataset's domain quality on the machine learning baseline, underscoring the need for further research into cross-domain machine learning applications.

Systematic review approaches are exemplified by the work of [16], who examined automatic question classification methods based on computer programming exams. Additionally, [17] reviewed NLP techniques and proposed strategies, including using lemmas instead of words, to enhance the Unified Medical Language System (UMLS). The study also suggested that medical students' medical documentation could benefit from a spell-checker enhanced by NLP, which would provide real-time educational feedback.

Furthermore, the review by [18] analysed techniques and algorithms for question classifications, revealing that SVM is the predominant machine-learning technique used for classification. The study identified bag of words (BOW) and term frequency-inverse document frequency (TF-IDF) as key feature extraction and selection techniques. The study verified the effectiveness of the BOW technique in response processing and identified SVM as one of the best algorithms for this type of problem. A systematic review by [19] focused on articles published between 2015 and 2019 related to auto-question generation. The study's conclusions emphasised the need for more extensive experimental reporting using standardised metrics and called for increased

research and evaluation of straightforward approaches.

Additionally, the study by Ferreira-Mello et al. [20] examined various techniques for educational text mining, highlighting NLP as the most effective tool for the education industry. However, the study noted that many reviewed articles prioritised outcomes over the process, resulting in accurate but lacking interpretation. The increasing volume of data generated by educational processes, as well as the pursuit of efficiency and quality, have spurred significant research in NLP and machine learning techniques. Most studies have adopted systematic and comprehensive approaches to investigate various educational issues related to NLP techniques. This study employs a well-structured methodological approach to review NLP techniques and their associated strengths and challenges in extracting educational data, particularly in curriculum and examinations.

This work's contribution to the body of knowledge lies in its analysis of a broader range of innovative publications, offering insights into the state-of-the-art of NLP in processing education data. Given the versatility of NLP across multiple fields and the varied techniques employed in other domains, this article may prove valuable beyond education. Drawing on an extensive literature review, we compile evidence that can guide NLP researchers in selecting the most suitable algorithms and techniques for NLP tasks, while also informing NLP-based application developers about the latest strengths and challenges in document analysis.

3. PROPOSED APPROACH

A. Data Sources

This review investigates NLP techniques and their associated strengths and challenges in processing educational data, specifically focusing on examinations, curriculum, and educational content. To gather relevant data, we accessed the following scientific repositories: Science Direct https://www.sciencedirect.com/ and Google Scholar https://scholar.google.com/. These repositories house numerous comprehensive studies published in a variety of journals. The subsequent section provides detailed insights into the algorithms used for data retrieval from these sources.

B. Search Query Strategies

Our search queries are constructed by combining keywords using Boolean operators. Figure 1 illustrates the search query generated from three sets of search terms, representing NLP techniques and the types of documents to be included in the retrieved articles. The first set of keywords ($K's1$) pertains to NLP syntactic structure features (SF_n), covering lexical and syntactic analysis features that encompass elements like sentence-splitting, morphological analysis, tokenization, phrase structure, stemming, parts-of-speech (POS), and other aspects related to syntax and grammar relationships. The second set of keywords ($K's2$) revolves around semantic features (SM_n), focusing on the meaning of words and their contextual relationships within sentences. Additionally, this set includes features for sophisticated semantic representation of text data, including topic generation and document classification. The third set of keywords ($K's3$) addresses the document type (DT_n), such as educational content, questions, or curriculum. We generated an exhaustive list of search queries using a range of NLP techniques, encompassing syntactic structure, semantic representation, and advanced document analysis. These queries produced a set of results (R_{1-n}) from various repositories. After multiple tests involving different Boolean configurations, this search query proved to be the most effective in accessing the chosen databases and producing relevant results.

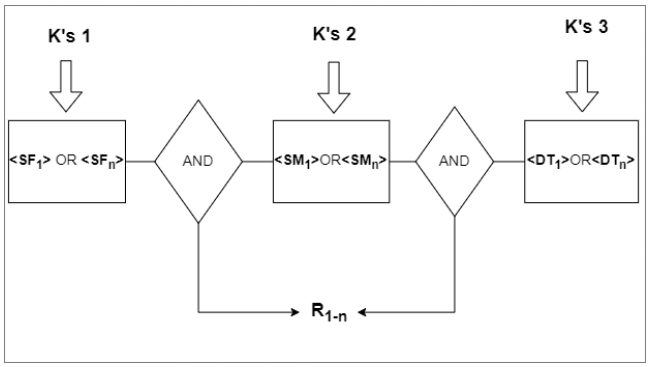

Figure 1. Flow diagram represents the combination of keywords in search queries

C. Procedures Used in Data Retrieval

The search query terms for Science Direct and Google Scholar Repositories, shown in Table I, were generated following the query generation process described above. These terms encompass various combinations of search terms, as outlined in the search query strategies section.

TABLE I. SEARCH QUERY TERMS USED FOR DATA RETRIEVAL

Repositories	Search Query Terms
Science Direct and Google Scholar	NLP AND "syntactic features" OR "lexical features" AND "semantic features" OR "semantic representation" AND "Question classification" OR curriculum OR educational contents

After performing these queries in both repositories, we acquired a significant amount of data. In the Science Direct repository, we filtered the results using the "articles type"tool, specifically picking "research articles" and removing irrelevant materials such as review articles and encyclopaedias. Furthermore, we refined the outcomes to specifically focus on the topic area of "computer science" and choose the most pertinent articles. We obtained a total of 1,961 articles from various journals, including Procedia Computer Science, Neurocomputing, Knowledge-Based

Systems, Information Processing and Management, Journal of Systems and Software, Future Generation Computer Systems, Procedia Technology, Decision Support Systems, Information and Management, and Applied Soft Computing. After removing 110 review articles, we obtained 1,564 papers from the Google Scholar platform. This left us with a total of 1,454 articles. In addition, we chose 10 more articles from relevant reference lists.

D. Inclusion Criteria

The retrieval algorithm provided a dataset of 3,425 articles from both repositories for further screening. These papers used the inclusion criteria presented in Table II. The inclusion criteria encompassed publication years, titles, language, abstract content, and specific keywords related to natural language processing approaches employed in the assessment of educational materials, such as exams, syllabi, educational texts, and curricula.

TABLE II. INCLUSION CRITERIA

SN	Factor	Inclusion Criteria
1	***Year***	2010 – 2023
2	***Language***	English
3	***Types of Publications***	Peer-reviewed working papers and books
4	***Title***	Relevant concept per study
5	***Abstract***	Keywords related to the study
6	***Text Screening***	NLP techniques in evaluating educational contents such as examination and curriculum

The title inclusion criteria initially excluded 3,156 articles, leaving us with 269. The second screening phase involved assessing keywords and abstracts, resulting in the exclusion of 172 articles. In the third screening phase, a detailed examination of the full text led to the removal of 26 articles containing unrelated educational data on examinations, curriculum, and educational content. As shown in Figure 2, 71 articles met the established criteria and were selected for inclusion in our study.

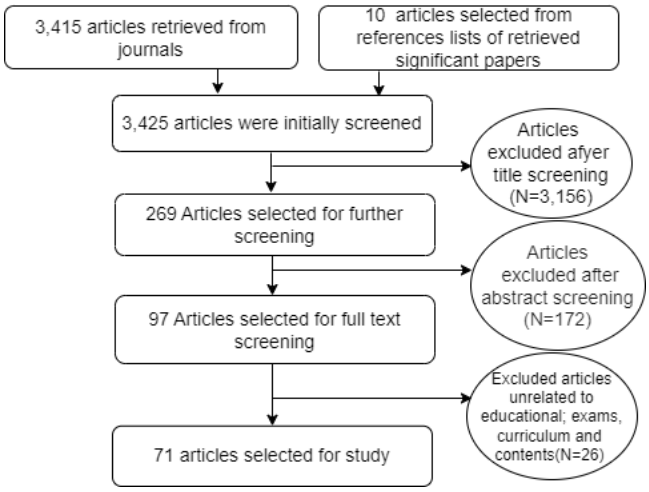

Figure 2. Inclusion and Exclusion Criteria

E. General Analysis of the selected Articles

We thoroughly reviewed the remaining 71 articles relevant to our study. We analysed these cases for similarities to enhance their presentation. Figure 3 provides an overview of the general analysis of the retrieved articles published between 2010 and 2023. The articles cover various techniques for evaluating exam questions and assessments, including the BT model, factors like answer categories, pattern matching, and syllabus coverage, as well as studies on curriculum and syllabus evaluation techniques and educational documents like lecture content. These articles describe techniques for evaluating educational documents to determine their quality according to various standards. Studies have focused more on evaluating examinations based on BT and other quality factors, rather than curriculum and other educationally related data. This could be due to the challenges of evaluating examination questions, as mentioned by [21], or to the fact that high-quality examinations play a pivotal role as the primary means for assessing acquired skills and learning outcomes

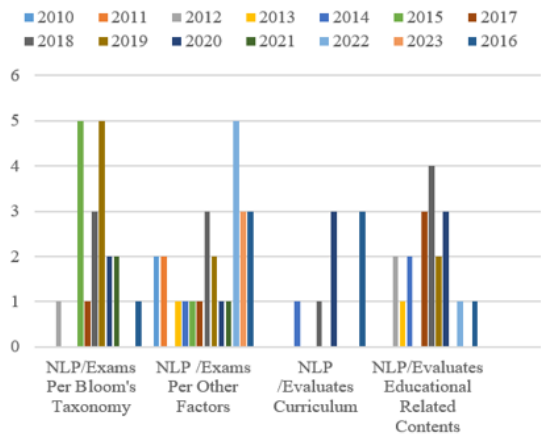

Figure 3. Analysis of the retrieved articles from 2010 to 2023

4. RESULT AND DISCUSSION

Across these articles, there are similarities in the phases involving text processing in NLP applications (Fig. 4). These phases encompass techniques for conversion, representation, dimension reduction, feature selection, feature extraction, and classification, all of which aim to process natural language, or text, into a format that machine learning can further utilise. These techniques vary in phases, as well as in their applications. However, most studies select their applicability based on task completion, minimal processing costs, semantic and syntactic performance portrayal, and efficiency.

The ability of NLP and machine learning to incorporate features that facilitate textual content comprehension makes them preferred for assessing various aspects of higher education quality. They use techniques rooted in natural language to establish an interpretative bridge between humans and machines [22]. Furthermore, NLP is enriched with numerous toolkits that enable the development of robust applications without starting from scratch [23]. Mathematical and statistical tools like BOW, TF-IDF, N-grams, and topic modelling are used with traditional or count-based features. For document analysis, SVM, Random Forests (RF), Decision Tree (J48), and Naive Bayes (NB) models are also used. Deep learning methods, such as Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU), as well as techniques like Continuous Bag of Words (CBOW), Word2vec, skip-gram, and others [24], fall into a different group. While deep learning often outperforms traditional methods, several hybrid approaches have been developed to enhance traditional techniques and boost their effectiveness [25].

A. NLP Techniques for Feature Engineering Based on Text Input Data Analysis

The reviewed literature addresses the application of NLP and machine learning in educational data analysis. These applications leverage machine learning algorithms to intelligently assess the quality of academic data, including examinations and curriculum [12]. NLP approaches are essential for mitigating the limitations of text mining. They entail preparing textual data and extracting relevant features, which improves textual information comprehension. Consequently, these techniques improve text mining algorithms and yield ideal outcomes for the assigned job [26]. A prominent illustration of NLP's application lies in document analysis, a fundamental process with wide-ranging implications across diverse fields, including extracting valuable insights from text and its application in numerous domains. Document analysis comprises three essential stages: preprocessing, document representation, and classification (as depicted in Figure 4).

The document analysis task, which is fundamental for activities such as extracting important insight from text and applications in numerous fields, is one example of how NLP may be used for automated text analysis. The document analysis processes and techniques are depicted in Figure 4. Text preprocessing involves a variety of techniques, including tokenization, lowercasing, and stemming, among others. These techniques are utilized to clean and transform text documents, preparing them for subsequent processes. Text representation techniques, including BOW, N-gram, and similar methods, convert text into a mathematical computational format often referred to as "feature extraction." This feature extraction is a crucial step in the classification process. Methods such as Support Vector Machines (SVM), Naive Bayes (NB), and others categorize the represented text for various tasks and applications in the classification process.

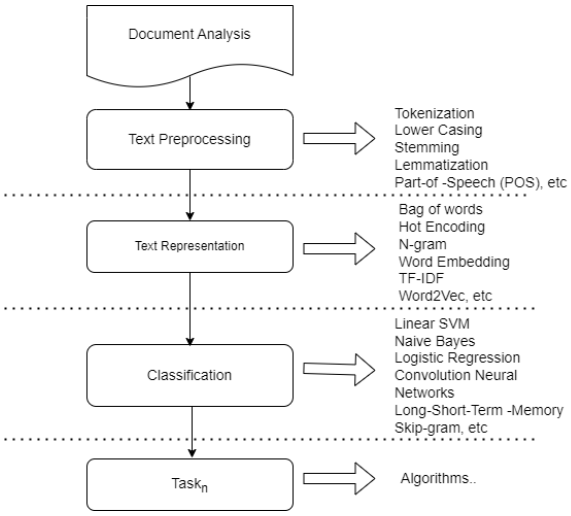

Figure 4. Text Analysis Processes and Techniques

NLP can accomplish tasks that include matching, classification, translation, structure prediction, and the sequential decision process [27]. These tasks require techniques for extracting features based on syntactic and semantic features and placing them in a format that can be used by the task at hand [28], i.e., matching or benchmarking. [22] describes and categorises the NLP techniques for feature engineering in three lines, which can be merged to construct a rich feature representation of text data. These are syntactic-phrase-based features, parse-tree-based features, entity relation features, pure statistical features, and latent semantic features. The first line includes syntactic-phrase-based and parse-tree-based features, which provide meaningful insights into sentence or phrase meanings [29]. Entity relation features aim to identify and characterise semantic relationships between text entities [30]. The second line comprises pure statistical features, such as statistical phrases, frequent word sets, and sequential word patterns. These features extract patterns of groups of words from the input text and can be used as features [26]. The third line includes techniques like singular value decomposition (SVD), probabilistic topic modelling, probabilistic latent semantic index, and latent

Dirichlet allocation (LDA). These techniques perform dimension reduction to represent concepts rather than raw terms, and they use probabilistic models to discover phrase co-occurrence patterns that correlate with semantic topics [31].

Moreover, we categorize NLP techniques for feature engineering into traditional and advanced models. Traditional or count-based feature engineering extracts features from text using mathematical and statistical methods like the BOW model, TF-IDF, N-grams, and topic modelling. Search engines, document clustering, and information retrieval commonly use these features to assess document similarities. However, traditional methods often neglect important aspects such as semantics, word structure, and context, potentially leading to suboptimal results[25].

Advanced feature engineering, on the other hand, leverages the limitations of traditional features to create vector representations of words. Advanced feature engineering includes predictive techniques such as neural network language models that analyze word sequences and predict words based on their context, known as embedding features (ref 25).

B. Overview of NLP Techniques for Examination Analysis

Examinations are the most common method for evaluating university students' cognitive capacity. Researchers have conducted several studies to automate the evaluation of examination quality. Some studies use BT [32], [33], [34], [35], Solo [36], construct question-answering systems [37], and match the examination with learning objectives [38], which are just a few of them. The classification process is critical for completing the tasks mentioned above; the next sections offer an overview of the techniques used in examination classification.

1) *Techniques for Preprocessing Examinations*

The preprocessing techniques primarily centre on tasks aimed at cleansing and transforming data into a suitable format for subsequent processes [35]. These techniques include tasks such as removing diagrams, symbols, descriptive text preceding the questions, and words with less than three letters, as well as removing punctuation and eliminating non-Unicode characters [38]. Additional techniques include converting characters to lowercase, removing punctuation marks and numbers, tokenization (breaking content into tokens), phrase segmentation, stop-word removal, part-of-speech (POS) tagging, and parsing. [39], [33], [36], [34], [40], [12], [41], [42], [38], [43]. Later processing stages address semantic and pragmatic aspects, while most of these techniques deal with the appearance of word forms using lexical and contextual information. Normalisation is another technique that reduces word form variability to a consistent style [12], [41], [44]. Some studies, like [43], [41], [42], [36], [12], and [44], used stemming. However, lemmatization is better for question classification because it finds valid root words that are important for getting semantic information from test questions [45].

Furthermore, Manjushree et al. [40] use tokenization techniques, which convert lowercase to uppercase, break sentences into words, and return frequently used phrases while excluding less common ones. Label encoding transforms labels into machine-readable forms, and pad sequence ensures that sequences in a list have the same length. Preliminary preprocessing is essential for preparing text input for subsequent machine-learning processes. Several challenges may arise from preprocessing steps, including removing critical symbols, which can lead to unintended consequences, as stated by Pintar et al. [42]. For example, the custom symbols introduced to prevent symbols like 'U' for voltage and 'I' for current from being eliminated as stop words could cause issues [42].

Lemmatization can also be hard for languages with few resources, like Swahili, because it might not establish a structured semantic relationship between words, which could lead to mistakes [46], [47]. For languages with lots of resources, like English, the WordNet lemmatizer is used to find real root words marches [36]. Constructing a structured semantic relationship between words in low-resource languages such as Swahili may be difficult, resulting in lower performance by misrepresenting the correct meaning of the word in question [46], [47]. In high-resource languages such as English, the WordNet lemmatizer is used to retrieve real root words via matches [36].

2) *Techniques for Text Representation*

Many studies do not differentiate significantly between preprocessing and text representation. Text representation encompasses techniques that convert text into numeric vectors, which can be semantically or syntactically evaluated [48]. This also includes dimensionality reduction via feature extraction, adding additional features to retain useful information, and feature selection, retaining a subset of the original features [45]. Feature extraction involves techniques such as unigrams, wh-words, word shapes, tagged unigrams, headwords, related word groups, hypernyms, and tagged bigrams [37], [49], [50]. Bag of Words (BOW) [49], verbs or keywords extraction [33], [36], [34], TF-IDF weighting followed by latent Dirichlet allocation (LDA) to find new topics based on stemmed words [38], N-grams or unigrams for word representation, and TF-IDF for counting verbs and nouns [39], [44] are some other methods.

Other techniques include enhanced E-TFIDF [41], modified TF-IDF to TFPOS-IDF [12], grammatical patterns that relate to the text's words [51], and bag-of-concepts [48]. The N-gram, also known as a unigram, shows words, while the term reference counts the number of verbs and nouns [39]. The word2vec embedding vector includes different ways to show text, such as the continuous bag of words (CBOW), which creates word representations by picking a centre word from a window of chosen context words, and

the skip-gram, which creates word vector representations by picking the context words around a given word [52]. Feature selection methods encompass techniques like latent semantic analysis using singular value decomposition, which uncovers hidden links between phrases and concepts [53], mutual information, which selects the most crucial features, chi-square statistics, which selects features strongly dependent on the response, and odd ratios for selecting appropriate feedback words [49].

Gobbo et al. [54] employ various methods to convey short responses' lexical and semantic characteristics in developing the automated short answer grading (ASAG) system. Combining these lexical traits with the collected embeddings results in a feature matrix that machine learning techniques can utilize. From a lexical standpoint, text data is represented through the utilisation of term frequency and inverse document frequency, thereby forming a bag-of-words approach. However, it is important to note that this approach has certain limitations, despite its impressive results. Notably, using regressors presents inherent disadvantages that impact the final score. The regression process can estimate values below the minimum threshold (zero) and above the maximum threshold (five) due to the absence of constraints. Wang et al. [55] introduce the semantic knowledge mapping network (S-KMN) to improve quiz question annotation. It combines semantic feature learning and knowledge mapping, addressing limitations in existing studies. Semantic feature extraction uses BERT, while a matrix-based vector facilitates knowledge feature extraction. However, the model's generalization and difficulty modelling word-question relationships hinder its performance in short question text and sparse semantics. Hamza et al. [56] introduced innovative methods for classifying Arabic questions, incorporating sentence transformer representations. The designers designed these methods to overcome the limitations of conventional techniques such as TF-IDF and word embeddings, especially in handling polysemous words. Their model, built upon Arabic bidirectional encoder representations from transformers (AraBERT), outperforms previous approaches in Arabic text classification, achieving an impressive accuracy rate of 94.19%. This research represents a substantial contribution to the field of Arabic language processing and classification.

The findings reveal both the strengths and limitations of textual representation techniques. Table VI highlights some salient features, including the strengths and limitations of NLP techniques. For example, common techniques like BOW have conceptual and contextual information in text, high dimensionality, and sparsity issues [48]. Furthermore, BOW fails to preserve important proximity information [57]. Another issue with the TF-IDF method is that it may give the same weight to different words, and some experiments don't cover all of Bloom's taxonomy levels [38]. Factors influencing technique performance include data size, where larger datasets maximize efficiency, and dimensionality, as low dimensionality improves performance by reducing computational costs and storage space [58]. The emergence of large language models like GPT, BERT, and their derivatives has greatly enhanced the capacity for text extraction and the determination of semantic and contextual meaning. However, due to the complexities of languages, extensive research is still required, especially in scenarios with limited linguistic resources.

3) *Approaches Based on Question Classification*

Several articles cover NLP techniques for question classification; the techniques employed in each category differ based on the criteria or goal at hand. Table III summarizes some common approaches to classifying and rating exams, as discussed by various authors.

TABLE III. QUESTIONS CLASSIFICATION SUMMARY BASED ON VARIOUS CRITERIA AND TECHNIQUES

Authors	Criteria	Evaluate attributes	Classification techniques
[59]	Pattern Matching	Categories questions into six functional	Multi-Layer Neural Network (MLN
[60]	Biomedical Questions	Classifies into three categories: YES/NO, Factoid, and Summary Questions	Algorithm-based question pattern classification
[51]	Syntactic Classification	Syntactic categorization into six categories	Support Vector Machine (SVM), Random forests (RF), Decision Tree (J48), Naïve Bayes (NB)

Authors	Criteria	Evaluate attributes	Classification techniques
[38]	Objectives-based Classification	Automatically label practice opportunities based on anticipated learning outcomes	Support Vector Machine and Extreme Learning Machine
[40], [35], [61], [62], [63]	Bloom's Taxonomy (BT)	It categorizes questions based on cognitive levels	SVM, NB, Logistic Regression, and Decision Trees, rule-based, RNN, LSTM
[64]	Syllabus Coverage	Examines the coverage of syllabus by question paper	Matrix Similarity via matrix representation vector an extension of VSM
[65]	Closed-Domain Question Answering	Assesses Overall performance and fine-grained analysis	Combined rule-based and machine-learning approaches

Studies classify question classification into rule-based, machine learning, hybrid, statistical, and deep learning approaches. These approaches employ feature extraction and selection methods for classifying questions based on expected metrics. Table IV presents the classification approaches, along with their strengths and limitations.

TABLE IV. CLASSIFICATION APPROACHES AND CHALLENGES FACED

Authors	Classification Approach	Strength	Limitations
[35], [33]	Rule-based approach	High accuracy but intensive demand of human effort	Time-consuming, tedious, lower performance compared to ML and less effective when classifying questions using Bloom's Taxonomy
[45], [65]	Machine Learning-based Approach	Enable the creation of superior question classification systems that utilize diverse features.	SVM model's high-dimensional outcome clarity issues and computational complexity; Naïve Bayes model's accuracy dependent on training data size
[66]	Statistical-based Approach	Suitable for large data but relies on conceptual understanding rather than semantic evaluation.	Reliant on conceptual understanding rather than semantic evaluation

Authors	Classification Approach	Strength	Limitations
[65], [67]	Data-based Approach	combine the strengths of various techniques while mitigating their limitations	Handling Persian inquiry classification and encountering English terms in Persian inquiries can be challenging and may require human intervention.
[40], [27]	Deep learning Approach	It supports automatic learning of multi-level feature representations with high precision but requires large data and incurs high computational costs.	Unsuitable for exam data due to data size requirements, expensive to train, and lack of theoretical grounding for selecting appropriate deep learning tools

Previously, techniques primarily relied on statistical and probabilistic approaches, emphasising a conceptual understanding of text. However, with the advent of NLP, the focus shifted towards semantic evaluation, enabling a more natural assessment of language likeness [66]. Statistical approaches include linear methods such as SVM, probabilistic topic models, and non-linear neural networks. Each has its strengths and limitations, and there are ongoing efforts to combine them to achieve optimal performance [68].

Additionally, some studies evaluate the performance of various methods, including [37], [49], and [50]. These studies assess classification performance by combining features extracted from lexical, syntactic, and semantic aspects. For example, [49] combines lexical features like unigrams and word shapes, syntactic features such as headwords and question categories, and semantic features like hypernyms. Similarly, [50] utilises lexical features like unigrams and bigrams, syntactic features like headwords, and semantic features like query expansion, question categories, and related words to create a feature known as question patterns. Table V illustrates that coarse-grained classification achieves better results by 1% in [49] and slightly lower results by 0.5% in fine-grained classification compared to [50]. However, coarse-grained classification falls short by less than 3% in [37]. According to [49], hypernym and question category features enhance the ability to identify relationships in naturally occurring text, improving performance in NLP tasks such as categorization.

TABLE V. COMPARE THE STUDY OF FEATURES, ALGORITHMS AND PERFORMANCE

Study	**Combinations**	**Classifier**	**Performance**	
			Coarse Grain	Fine Grain
[49]	U+H+HY+ WS+QC	Linear SVM	96.2%	91.1%
[50]	U+B+WS+H+ R+QE+QC+QP	Linear SVM	95.2%	91.6%

Most early approaches to question classification relied on rule-based procedures involving manually creating classification rules [69]. In their work, [69] employed a rule-based approach to assess the performance of various features. They used a rule-based question classifier to manipulate and generate features, which they then combined with other features within SVM to improve performance. Their study indicated that while headword features (H) alone yielded subpar results, combining them with category features (C) for coarse granularity improved results. Furthermore, combined with unigrams (U), the classifier achieved its best results [**?**]. In contrast, other studies focused on classifying questions according to Bloom's Taxonomy. They employed various techniques, such as preprocessing operations with a rule-based strategy for query categorization. This approach involved assigning weighted categories to overlapping keywords [35]. However, due to variations in background knowledge across domains, this technique sometimes results in inconsistencies and suboptimal classification performance. Another study modified the TF-IDF by introducing enhanced TFIDF (E-TFIDF). This enhancement gave more weight to verbs, nouns, and adverbs than other words. The results were then looked at with SVM, NB, and KNN classifiers, which showed that the improved E-TFIDF worked better than the others [41]. A different study compared the performance of three features—TF-IDF, TFPOS-IDF, and W2VTFPOS-IDF. The results showed that accuracy went up when traditional TF-IDF was changed to TFPOS-IDF, verbs were given more weight, and W2VTFPOS-IDF was used for context and high-quality feature vector representation. Specifically, the average accuracy with different classifiers was as follows: Logistic Regression and Support Vector Machine achieved 71.1%, 82.3%, and 83.7%, respectively, while for 600 questions within the same classifiers, the accuracy rose to 85.4%, 89.4%, and 89.7%. To put test questions into Bloom's Taxonomy, the authors in [44] used both syntactic and semantic information, such as part-of-speech tagging (POS)

and WordNet and the Lest algorithm. They represented words using N-grams or unigrams and employed SVM, NB, and J48 classifiers. The results showed that classifiers that combined multiple features outperformed those that did not, achieving higher f-measure scores, with SVM ranking as the top-performing classifier. The authors in [70] used techniques involving verbs extracted from sentences and stemmed using the Lancaster stemmer. The authors employed POS tagging to generate sentence skeletons and used WordNet to identify the correct word root. This approach resulted in an accuracy rate of 72.9%. The authors in [39] used TF-IDF and NB to classify examinations based on Bloom's Taxonomy cognitive levels. The authors applied preprocessing techniques such as dataset labelling, tokenization, stemming, and filtering. Additionally, feature extraction involved using the TF-IDF technique on a series of words, characters, and N-grams. The N-gram approach achieved the highest accuracy and precision at 85%. The mid-term and final exams from Telkom University's Department of Information Systems made up the dataset for this study.

Furthermore, several studies aimed to classify questions based on answering systems. [71] employed SVM as a classifier and a set of low-dimensional lexical and syntactic features to summarise the content of a larger set and classify questions. The obtained accuracy rates were 89.2% for course classes and 82.4% for fine classes. However, these results were slightly lower than those reported by [49] and [50]. This difference may be attributed to variations in the training dataset. It is noteworthy that the [71] study did not take into account semantic features. The other approach [51] classified questions using grammatical structure, syntactic features, and other techniques, as shown in Table IV. The J48 decision tree classifier outperformed other classifiers with an accuracy rate of 91.1%. Meanwhile, [72] applied four deep-learning approaches for question classification: CNN, GRU, LSTM, and CNN-GRU. They trained on 5,400 questions using Word2Vec embedding vectors like Skip-gram and CBOW and tested on 600 questions. Word2Vec effectively captured semantic and syntactic relationships between words in documents, enhancing classification model performance. In the skip-gram mode, CNN-LSTM and CNN-GRU outperformed CBOW, achieving an accuracy rate of 93.7%. It is important to note that the English dataset reached 94.4% accuracy using the LSTM technique, while the Turkish dataset did not, likely due to its complexity. In the [73] study, on the other hand, they focused on feature extraction for medical question classification and added RNN, LSTM, and GRU to get more contextual information than the usual multilayer perceptron. These conventional models treat input data as independent, limiting their ability to capture contextual information. The score was 54% lower than the traditional GRU, possibly due to the RNN model's interference with the extraction of final features. In reference [74], the authors employ traditional and modern NLP techniques to extract information from textual responses, specifically focusing on detecting incoherence in open-ended responses to inquiries. The study investigates three categories of models: ensemble, deep, and shallow. Shallow models involve SVMs, NB, and feature extraction techniques like TF-IDF and word embeddings. Feature extraction encompasses BETO, a Spanish equivalent of BERT, hand-engineered features, and word embeddings. Deep learning models leverage transformers and BETO for representation learning. The results highlight the effectiveness of the multi-layer neural network and deep learning model, achieving an accuracy of 79.15%, particularly in identifying incoherence. A study by [75] uses a multi-task model for query understanding (MTQU) to make named entity recognition and question classification better in systems that answer questions. The model has five layers: bidirectional long short-term memory (BiLSTM), attention, pooling, and output. It uses morphological analysis, feature representation, lexical elements, syntactic variables, BiLSTM, attention, pooling, and output to deal with the lack of data in Kazakh question comprehension. The model outperforms prior models, with QC accuracy at 92.28, NER F1 score at 91.73, and sentence-level semantic frame accuracy at 83.58. The model's multi-feature input layer directly integrates QC and NER tasks, enabling concurrent progress. The study [76] describes a complete model for classifying Arabic test questions. It has five parts: M-TF-IDF, ArELMo, BiGRU, CNN, and a mechanism for paying attention. M-TF-IDF processes keywords, ArELMo represents word vectors, and Bi-GRU analyzes forward and backward Arabic questions, extracting contextual and semantic features. The attention mechanism enhances keyword value for CNN feature extraction. M-TF-IDF outperforms TF-POS-IDF with recall at 84.32, precision at 85.84, and accuracy at 84.26. The modified BiGRU-CNN model also outperforms the LSTM, LSTM-CNN, and BiGRU models. Larger models such as BERT and GPT require further exploration. Here, [77] tests how well pre-trained embeddings from transfer learning models like embeddings from language models (ELMo), BERT, generative pre-trained transformer (GPT), and GPT-2 work in automatic short answer grading (ASAG). We compare this evaluation with prior methods that use concept mapping, facet mapping, and conventional word embeddings for semantic feature extraction. The dataset includes 2,273 answers from 31 students, responding to 80 questions across 10 assignments and 2 exams. The sole text preprocessing technique is tokenization, assigning pre-trained embeddings to tokens for cosine similarity calculation. ELMo outperforms baseline models with a Pearson correlation coefficient of 0.485, largely attributed to its extensive domain data. The study shows that more research needs to be done on different sentence embedding methods because the current method relies on Sum of Word embeddings (SOWE) in a high-dimensional hypothesis space. [78] uses a Siamese-stacked bidirectional long-short-term memory model to identify semantic textual similarities between student and model answers. The model processes both inputs simultaneously, capturing complex contextual information and extracting features. The recommended architecture for the data struc-

tures course dataset provides the best Pearson correlation value of 0.668. However, the study struggles to identify brief replies, only including one to three words, despite its substantial findings on both domain-specific embedding and stacked BiLSTM networks. The study was led by [79] and aims to give accurate answers to "why"-type questions that aren't factoids by looking at lexical-syntactic, semantic, and contextual factors using deep learning frameworks. The approach uses an ensemble ExtraTreesClassifier to re-rank answer candidates based on importance scores, achieving a mean reciprocal rank (MRR) of 0.64. We perform answer validation by matching answer types, with semantic features being the most crucial factor. The study shows that semantic characteristics are the most important factor, achieving a mean reciprocal rank (MRR) of 0.64. However, there is a gap in incorporating discourse processing and commonsense reasoning, suggesting potential areas for improvement and application in restricted domain question answering systems. The study highlights the need for further development and application in these areas.

The study by [80] developed an automated system for creating factual-based questions using NLP methods like syntactic and semantic feature extraction, paraphrasing, and evaluation metrics. The system uses a rule-based approach, but its challenges include question quality assessment, syntactic complexity management, and question correctness. The system needs to improve its handling of linguistic subtleties and paraphrase strategies.

The author [81] aims to improve question classification (QC) by utilising data augmentation techniques to create more training examples, reduce the need for large datasets, and address limited labelled data. The research uses NLP techniques to achieve state-of-the-art performance with fewer labelled cases, addressing the gap in expensive and time-consuming large labelled datasets.

C. Overview of the Educational Document Analysis and Incorporation Tasks

This review uses educational documents to represent course material, syllabi, and curriculum. Document analysis techniques are similar to question classification; however, the research reveals that question categorization faces the same challenges compared to document classification due to the shorter length of questions. As a result, these two domains require distinct approaches [21]. NLP research on educational document analysis encompasses studies identifying learning concepts from learning resources [82], [83]. For instance,[48] introduced the bag-of-concepts model to address traditional bag-of-words (BoW) issues in document classification tasks, particularly in the text representation process. This model aimed to reduce dimensionality and tackle sparsity concerns.

Additionally, the study by [81] evaluates the similarity of syllabi among higher education institutions using the UNESCO knowledge area classification. The authors [81] automatically relate the topic to the course book and check for missing parts in the course specifications, while [82] identify the course's knowledge performance indicators. The study by [83] deployed NLP rules to locate specific and relevant opinion words about which feedback is given, as well as the opinion's orientation, i.e., positive, negative, or neural. The authors [84] evaluate teaching material and assessment based on learning outcomes, while [85] evaluate the coherence of an academic curriculum. Furthermore, the study by [86] implemented multi-sentence classification on a large number of documents using CNN.

In summary, NLP-based educational document analysis encompasses a broad spectrum of tasks, whereas examination analysis studies predominantly revolve around question classification using Bloom's Taxonomy. This distinction underscores the diversity and complexity of tasks within the two domains.

1) *NLP Techniques for Content Analysis on Educational Documents*

Content analysis in educational documents involves a sequence of techniques and approaches (Fig. 4). In [83], the study employed preprocessing techniques that included removing mathematical symbols, variables, and numbers in context, parsing sentences and paragraphs, eliminating punctuation marks and special characters, converting characters to lowercase, and removing white spaces. N-grams and TF-IDF were then used to extract features for SVM training to determine whether a document contained a learning concept. Furthermore, dimension reduction techniques like singular value decomposition (SVD), principal component analysis (PCA), and multi-dimensional scaling (MDS) were applied to reduce feature space noise. However, the application of dimension reduction strategies resulted in a decrease in the system's accuracy, suggesting the omission of many learning concepts. Dealing with unbalanced data also affected the performance of SVM. In contrast, [82] aimed to identify core concepts in educational resources by assessing their relevance to domain topics. For dimensionality reduction, the authors used SVM classifiers, preprocessing techniques, TF/IDF, and LSA, as well as TF/IDF for feature selection. They used latent semantic analysis to find important features and compared 1,442 course outlines from computer science classes at higher education institutions in Ecuador, putting them into groups based on UNESCO knowledge areas. However, challenges arose due to the high similarity between courses with different contents but similar topics. The authors in [84] used SVM classifiers, preprocessing techniques, TF/IDF and LSA for dimensionality reduction, and TF/IDF for feature selection. The authors then utilize latent semantic analysis to identify relevant features, and employ cosine similarity methods to assess the similarity of 1,442 computer science course syllabi from Ecuadorian Higher Education Institutes, basing the classification on UNESCO knowledge areas. However, the study by [67] encountered challenges, including a high degree of similarity in courses with different

contents but with the same topics. [85] used text mining techniques to identify a course's Knowledge Performance Indicators (KPIs) through tokenization, stop word removal, stemming, and extracting synonyms and keywords representing intended learning outcomes (ILOs). Term frequency was calculated for knowledge, understanding, intellectual, professional, and practical skills. The authors in [86] use text mining to find the course's Knowledge Performance Indicators (KPIs). They do this by preprocessing the text using techniques like stemming, tokenization, and stop word removal. They also find synonyms for words and keywords that show the intended learning outcomes (ILOs) and then figure out the frequency of terms based on intellectual skills, practical skills, knowledge and understanding, and term frequency. [87] evaluated teachers and courses, extracting general topics using supervised machine learning and employing NLP techniques to identify specific opinion terms and their associated feedback. The techniques included Apache OpenNLP for preprocessing, TF-IDF in String2WordVector for feature extraction, and Naive Bayes multinomial classification for text classification. The resulting processes achieved recall and precision rates of 83% and 84%, respectively, with limitations including the system's inability to handle new input words, incorrect English words, and potential misassignment of polarity to non-existent words in SentiWordNet. In another study, [88] classified academic and professional counseling queries into categories based on Holland's RIASEC topology. RIASEC represents six personalities—realistic (R), investigative (I), artistic (A), social (S), entrepreneurial (E), and conventional (C)—corresponding to career choices or educational program environments. Multiclass neural networks outperformed SVM and feature-based classification algorithms, allowing the processing of datasets with complex biomedical issues. Preprocessing and feature engineering models were employed, converting raw text to integers for input into the multiclass neural network. The dataset was split into 70% for training and 30% for testing, with the multiclass neural network achieving superior classification results. To assess teaching materials and assessments based on learning outcomes, [89] employed four components. These components included analyzing learning outcomes and levels in Revised Bloom's Taxonomy, assessing lecture material fairness, assessing question paper fairness, and assessing practical session fairness. The dataset consisted of 600 learning outcomes from various modules and underwent preprocessing, including tokenization, stop word removal, text conversion to lowercase, and the Natural Language Toolkit (NLTK) application. A recurrent neural network with LSTM architecture was employed to classify module lessons and learning outcomes, summarise lecture PowerPoint presentations, and review slides covering learning outcomes. Additionally, [90] examined three models—CNN, standard LDA, and modified LDA with TF-IDF—for multiple text classification on a large document. The improved LDA improves accuracy from 60% to 74.44% and reduces time from 4.04 to 3.02. Nevertheless, CNN outperformed both LDA models with an accuracy of 94.7%. Preprocessing techniques included tokenization, removal of words with less than three characters, stop-word elimination, lemmatization, and stemming. [24] introduced a text similarity approach that considered the semantic sequence of words rather than syntax. Preprocessing involved tokenization, lowercasing, and stemming of short text. The method used a mix of dependency parsing and lexicon embedding connected to outside sources like ConceptNet to check the coherence of words. Sentence pair similarities were calculated using bag-of-words (BoW) vectors. Then the sentence pairs' similarities were calculated using the bag-of-words (BoW) vector. [48] identified entities and concepts in documents through preprocessing techniques such as sentence segmentation, word tokenization, and POS tagging. Entity sense disambiguation was employed to address polysemous terms and enhance text representation. The study introduced the concept of bag-of-concepts to address the limitations of traditional BoW models in document classification tasks, particularly dimensionality expansion and sparsity. However, according to [24], deep learning provides a longer vector of text representation that contains expanded text like sentences and paragraphs, resulting in a more efficient form of text presentation and hence increased text or document classification accuracy. [91] evaluated the coherence of an academic curriculum using datasets that included Database Design Concepts (179), Data Mining (212), Business Process Management (468), and Network Security (156). Preprocessing techniques removed syntactic variations, such as plurals and capitalization, replaced synonyms, abstracted specific concepts, and eliminated index terms with fewer than two occurrences. POS identification during preprocessing revealed that noun extraction conveyed the most relevant meaning, enhancing contextual entailment. However, the techniques employed for dimensionality reduction were inadequate for reducing processed data and minimising computational costs. Vo et al. [92] used named entity recognition (NER) and a hybrid course recommendation system to categorise text into predefined groups. The CSIT-NER model (Computer Science and Information Technology), trained on StackOverflow and GitHub, extracts tech-related details automatically. The hybrid recommendation system integrates data from various sources, including job websites and online platforms, to offer personalised course suggestions. Data annotation involves manual or automated labelling to train the CSIT-NER model, enhancing its ability to provide accurate recommendations. However, the study lacks a detailed discussion on potential limitations or challenges in the data annotation process. The author's [93] examines SBERT, ADA-002, and ConceptNet embeddings alongside knowledge graph embeddings for educational content recommendation, noting benefits like unsupervised linkage and semantic similarity. It suggests personalized adaptive systems for quizzes.

TABLE VI. CRITICAL SALIENT FEATURES INTRODUCED IN ARTICLES

Year & Authors	Topic Covered	Data Used	Used Techniques	Strength of the techniques	Limitations of the techniques
[37]	Examination Questions Classification	6000 questions from the University of Illinois Urbana Champaign (UIUC) were mapped to the question's taxonomy	Unigram, wh-words, related word group, word shapes, tagged unigram, bigram, headwords, hypernyms	Lower computational cost compared to the state of the art at the time	A challenge in establishing the one-to-one relationship between features and analysis levels [94].
[35]	Automated Analysis of Exams based on Bloom's Taxonomy	70 programming exam questions (training), 30 (testing)	Preprocessing, stop words removal, stemming, lemmatization, POS tagging, and A rule-based approach are utilized to discover important keywords and verbs that determine the question's categorization.	Categorized the questions per cognitive level	Potential inconsistency due to overlapping keywords in Bloom's taxonomy
[49]	Question Classification using Semantic, Syntactic, and Lexical Features	Varies from 1,000 to 5,500 questions	Preprocessing, stemming and stop word removal. Text representation: bag-of-words, unigram, Headword, Hypernyms, Word shapes, Questions Category. Using Linear SVM classifier	Improved performance in categorizing questions	The heavy computational cost for syntax feature extraction [95], high dimensionality, and sparsity issues [48], also, fail to preserve the necessary proximity information as the number of unique words grows [57].
[71]	Classification of questions and large information using SVMs, forward-selection algorithm, based on newly introduced features	5,452 training questions, 500 testing questions	Preprocessing, tokenization, tagging, stemming, and parsing. Text Representation Techniques: Unigrams, Principal Wh-Word, Bigrams, Head-Word, Head-Verb and Multiple-Head-WoRDS	reducing the overall number of features i.e., semantic features hence reducing dimension,	Poor discrimination for some Principal-Wh-Words, classifier reliance on training data

Year & Authors	Topic Covered	Data Used	Used Techniques	Strength of the techniques	Limitations of the techniques
[43]	Classification of Exam Questions Based on Linguistically-Motivated Features	600 questions from Najran University	Preprocessing: tokenization, POS, Stemming, Lemmating, N-gram. Text representation: Unigrams, Bigrams, Trigrams, POS Bigrams, POS Trigrams, Word/POS Pairs, and Stem N-grams. Several classifiers were trained and tested: NB, Logistic Regression, SVM, Decision Trees	The use of SVM and Logistic regression, combined with the addition of N-grams like bigrams and trigrams, significantly improves performance by maintaining local word sequence ordering.	Lack of linkage between feature taxonomy and analysis level, neglect of semantic structure of questions
[82]	Identifying Core Concepts from Educational Resources	340 sentences from digital resources	Preprocessing, lemmatization, stopwords removal, string similarity, semantic similarity, generative model, shallow features	evaluate the document's core concept in terms of how well it embodies basic concepts from related subject domains.	Frequent use of shallow or weak semantic/textual features for semantic similarity computation
[36]	Classify exam questions by developing rules that identify categories and assign weights based on Bloom's Taxonomy	85 exam questions (training), 62 (testing) from Moratuwa University	Tokenization, lemmatization, POS tagging, rule-based approach with lemma similarity	The preprocessing phases comprise the selected tools that provide the appropriate and accurate format of text for the next phases	High lemma similarity for specific question verbs, potential inaccuracy in classifying all words as verbs
[38]	Automatic Labeling of Course Questions for Alignment with Learning Outcomes	150 questions from an electrical and electronic engineering course	Preprocessing, two phases, TF-IDF combined with ELM	Flexibility in training and testing datasets, good performance compared to traditional techniques like SVMFlexibility in training and testing datasets, good performance compared to traditional techniques like SVM	Limitations in evaluating multi-domain levels due to collapsing six levels into three, differing word weight nomenclature for the same sentences

Year & Authors	Topic Covered	Data Used	Used Techniques	Strength of the techniques	Limitations of the techniques
[34]	Bloom's Taxonomy and Rule-Based Question Analysis for Assessing Exam Papers	Over 900 short essay questions from 30 papers at Sabaragamuwa University	Preprocessing, tokenization, removal of non-letter characters, rule-based categorization	Effective assessment of question quality according to Bloom's Taxonomy	It requires significant manual labor, such as adding rules to each category.
[48]	Document Classification using the Bag-of-Concepts Model	Approximately 1,503,803 papers from various sources	Preprocessing, LSA and LDA-based dimension reduction, word2vec and Doc2vec for text representation	Capturing semantic and conceptual information is essential for document classification	Limited to concept and word level, lacks sentence-level semantic understanding
[91]	Evaluating an academic curriculum's coherence	The course and number of concepts include; Database Design Concepts (179), Data Mining (212), Business Process Management (468) and Network Security (156) are the dataset used.	Preprocessing techniques include removing syntactic variations like plurals and capitalization, then replacing synonyms, reducing idea space by abstracting certain extremely particular concepts, and removing index terms with less than two occurrences and POS identification	In the POS process, noun extraction techniques convey the phrase's most relevant meaning, thereby increasing contextual entailment.	The techniques utilized for dimensionality reduction are insufficient to reduce the processed data and thereby minimize computational costs.

5. CONCLUSION

In summary, NLP processes for text analysis share common elements, including preprocessing, text representation, classification, and algorithm application tailored to specific tasks. Many studies adapt existing techniques or introduce innovative methodologies to enhance performance. Nevertheless, they face challenges, such as limitations in representing local aspects and aligning with local educational standards. There is a clear need for further research, particularly focusing on local aspects, low-resource languages, semantics, and contextual nuances in terminology. This study highlights that techniques for analyzing and evaluating educational data have strengths and limitations, evident in aspects like local context, benchmarking against regulatory criteria, sentence-level semantic comprehension, and low-resource languages. Consequently, further investigation into NLP techniques is recommended, given the unique characteristics of educational content, which can have different implications at both local and global levels. Moreover, there is a need for practical exploration of domain-specific feature extraction and classification for quality aspects in academic data, including industrial skills, and further examination of the impact of hyperparameters on training data. Additionally, while there has been substantial focus on feature extraction in examinations, there remains limited research on curriculum, syllabus, and academic content, emphasizing the importance of exploring features for these vital documents linking education with industrial skills.

REFERENCES

[1] P. M. Mbithi, J. S. Mbau, N. J. Muthama, H. Inyega, and J. M. Kalai, "Higher education and skills development in africa: An analytical paper on the role of higher learning institutions on sustainable development," *Journal of Sustainability, Environment and Peace*, vol. 4, no. 2, pp. 58–73, 2021.

[2] M. Martin *et al.*, *Internal Quality Assurance: Enhancing higher education quality and graduate employability*. UNESCO, 2018.

[3] K. H. Mok and J. Jiang, "Massification of higher education and challenges for graduate employment and social mobility: East asian experiences and sociological reflections," *International Journal of Educational Development*, vol. 63, pp. 44–51, 2018.

[4] A. Leicht, J. Heiss, and W. J. Byun, *Issues and trends in education for sustainable development*. UNESCO publishing, 2018, vol. 5.

[5] R. Laurie, Y. Nonoyama-Tarumi, R. Mckeown, and C. Hopkins, "Contributions of education for sustainable development (esd) to quality education: A synthesis of research," *Journal of Education for Sustainable development*, vol. 10, no. 2, pp. 226–242, 2016.

[6] R. Bojorque and F. Pesántez-Avilés, "Academic quality management system audit using artificial intelligence techniques," in *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2019 International Conference on Human Factors in Artificial Intelligence and Social Computing, the AHFE International Conference on Human Factors, Software, Service and Systems Engineering, and the AHFE International Conference of Human Factors in Energy, July 24-28, 2019, Washington DC, USA 10*. Springer, 2020, pp. 275–283.

[7] S. K. Gill, A. Dhir, G. Singh, and D. Vrontis, "Transformative quality in higher education institutions (heis): Conceptualisation, scale development and validation," *Journal of Business Research*, vol. 138, pp. 275–286, 2022.

[8] G. A. Olcay and M. Bulu, "Is measuring the knowledge creation of universities possible?: A review of university rankings," *Technological Forecasting and Social Change*, vol. 123, pp. 153–160, 2017.

[9] A. S. M. Sohel-Uz-Zaman *et al.*, "Implementing total quality management in education: Compatibility and challenges," *Open Journal of Social Sciences*, vol. 4, no. 11, p. 207, 2016.

[10] K. Kawintiranon, P. Vateekul, A. Suchato, and P. Punyabukkana, "Understanding knowledge areas in curriculum through text mining from course materials," in *2016 IEEE international conference on teaching, assessment, and learning for engineering (TALE)*. IEEE, 2016, pp. 161–168.

[11] K. Sanvitha Kasthuriarachchi, S. Liyanage, and C. M. Bhatt, "A data mining approach to identify the factors affecting the academic success of tertiary students in sri lanka," *Software Data Engineering for Network eLearning Environments: Analytics and Awareness Learning Services*, pp. 179–197, 2018.

[12] M. Mohammed and N. Omar, "Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec," *PloS one*, vol. 15, no. 3, p. e0230442, 2020.

[13] J. West, "Validating curriculum development using text mining," *The Curriculum Journal*, vol. 28, no. 3, pp. 389–402, 2017.

[14] A. Sangodiah, R. Ahmad, and W. F. W. Ahmad, "A review in feature extraction approach in question classification using support vector machine," in *2014 IEEE international conference on control system, computing and engineering (ICCSCE 2014)*. IEEE, 2014, pp. 536–541.

[15] I. Thalib, I. Soesanti *et al.*, "A review on question analysis, document retrieval and answer extraction method in question answering system," in *2020 International Conference on Smart Technology and Applications (ICoSTA)*. IEEE, 2020, pp. 1–5.

[16] M. K. TAQI and R. ALI, "Automatic question classification models for computer programming examination: A systematic literature review." *Journal of Theoretical & Applied Information Technology*, vol. 93, no. 2, 2016.

[17] M. Chary, S. Parikh, A. F. Manini, E. W. Boyer, and M. Radeos, "A review of natural language processing in medical education," *Western Journal of Emergency Medicine*, vol. 20, no. 1, p. 78, 2019.

[18] V. A. Silva, I. I. Bittencourt, and J. C. Maldonado, "Automatic question classifiers: A systematic review," *IEEE Transactions on Learning Technologies*, vol. 12, no. 4, pp. 485–502, 2018.

[19] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, pp. 121–204, 2020.

[20] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 6, p. e1332, 2019.

[21] A. Sangodiah, R. Ahmad, and W. F. WAN AHMAD, "Taxonomy

based features in question classification using support vector machine." *Journal of Theoretical & Applied Information Technology*, vol. 95, no. 12, 2017.

[22] G. Dong and H. Liu, *Feature engineering for machine learning and data analytics*. CRC press, 2018.

[23] A. Pinto, H. Gonçalo Oliveira, and A. Oliveira Alves, "Comparing the performance of different nlp toolkits in formal and social media text," in *5th Symposium on Languages, Applications and Technologies (SLATE'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[24] E. Inan, "Simit: a text similarity method using lexicon and dependency representations," *New Generation Computing*, vol. 38, no. 3, pp. 509–530, 2020.

[25] D. Sarkar, *Text analytics with Python: a practitioner's guide to natural language processing*. Springer, 2019.

[26] C. Geigle, Q. Mei, and C. Zhai, "Feature engineering for text data," in *Feature engineering for machine learning and data analytics*. CRC Press, 2018, pp. 15–54.

[27] H. Li, "Deep learning for natural language processing: advantages and challenges," *National Science Review*, vol. 5, no. 1, pp. 24–26, 2018.

[28] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[29] W.-J. Huang and C.-L. Liu, "Exploring lexical, syntactic, and semantic features for chinese textual entailment in ntcir rite evaluation tasks," *Soft Computing*, vol. 21, pp. 311–330, 2017.

[30] J. Zhang and N. M. El-Gohary, "Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking," *Journal of Computing in Civil Engineering*, vol. 30, no. 2, p. 04015014, 2016.

[31] S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha, "Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond," *Mining text data*, pp. 129–161, 2012.

[32] D. A. Abduljabbar and N. Omar, "Exam questions classification based on bloom's taxonomy cognitive level using classifiers combination," *Journal of Theoretical and Applied Information Technology*, vol. 78, no. 3, p. 447, 2015.

[33] S. Dhainje, R. Chatur, K. Borse, and V. Bhamare, "An automatic question paper generation: using bloom's taxonomy," 2018.

[34] B. Kumara, A. Brahmana, and I. Paik, "Bloom's taxonomy and rules based question analysis approach for measuring the quality of examination papers," *International Journal of Knowledge Engineering*, vol. 5, no. 1, pp. 2–6, 2019.

[35] N. Omar, S. S. Haris, R. Hassan, H. Arshad, M. Rahmat, N. F. A. Zainal, and R. Zulkifli, "Automated analysis of exam questions according to bloom's taxonomy," *Procedia-Social and Behavioral Sciences*, vol. 59, pp. 297–303, 2012.

[36] K. Jayakodi, M. Bandara, and D. Meedeniya, "An automatic classifier for exam questions with wordnet and cosine similarity," in *2016 Moratuwa engineering research conference (MERCon)*. IEEE, 2016, pp. 12–17.

[37] B. Loni, G. Van Tulder, P. Wiggers, D. M. Tax, and M. Loog, "Question classification by weighted combination of lexical, syntactic and semantic features," in *Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings 14*. Springer, 2011, pp. 243–250.

[38] S. Supraja, K. Hartman, S. Tatinati, and A. W. Khong, "Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes." *International Educational Data Mining Society*, 2017.

[39] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text mining approach using tf-idf and naive bayes for classification of exam questions based on cognitive level of bloom's taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2019, pp. 112–117.

[40] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Classifications of the summative assessment for revised blooms taxonomy by using deep learning," *arXiv preprint arXiv:2104.08819*, 2021.

[41] M. Mohammed and N. Omar, "Question classification based on bloom's taxonomy using enhanced tf-idf," *Int J Adv Sci Eng Inf Technol*, vol. 8, pp. 1679–1685, 2018.

[42] D. Pintar, D. Begušić, F. Škopljanac-Mačina, and M. Vranić, "Automatic extraction of learning concepts from exam query repositories," *Journal of Communications Software and Systems*, vol. 14, no. 4, pp. 312–319, 2018.

[43] A. Osman and A. Yahya, "Classifications of exam questions using linguistically-motivated features: a case study based on bloom's taxonomy," in *The Sixth International Arab Conference on Quality Assurance in Higher Education (IACQA'2016)*, vol. 467, 2016, p. 474.

[44] O. J. Mohamed, N. A. Zakar, and B. Alshaikhdeeb, "A combination method of syntactic and semantic approaches for classifying examination questions into bloom's taxonomy cognitive," *Journal of Engineering Science and Technology*, vol. 14, no. 2, pp. 935–950, 2019.

[45] K. Makhlouf, L. Amouri, N. Chaabane, and E.-H. Nahla, "Exam questions classification based on bloom's taxonomy: Approaches and techniques," in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*. IEEE, 2020, pp. 1–6.

[46] B. Masua and N. Masasi, "Enhancing text pre-processing for swahili language: Datasets for common swahili stop-words, slangs and typos with equivalent proper words," *Data in Brief*, vol. 33, p. 106517, 2020.

[47] C. S. Shikali and R. Mokhosi, "Enhancing african low-resource languages: Swahili data for language modelling," *Data in brief*, vol. 31, p. 105951, 2020.

[48] K. Kitto, N. Sarathy, A. Gromov, M. Liu, K. Musial, and S. Buckingham Shum, "Towards skills-based curriculum analytics: Can we automate the recognition of prior learning?" in *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020, pp. 171–180.

[49] M. Mishra, V. K. Mishra, and H. Sharma, "Question classification

using semantic, syntactic and lexical features," *International Journal of Web & Semantic Technology*, vol. 4, no. 3, p. 39, 2013.

[50] N. Van-Tu and L. Anh-Cuong, "Improving question classification by feature extraction and selection," *Indian Journal of Science and Technology*, vol. 9, no. 17, pp. 1–8, 2016.

[51] A. Mohasseb, M. Bader-El-Den, and M. Cocea, "Question categorization and classification using grammar based approach," *Information Processing & Management*, vol. 54, no. 6, pp. 1228–1243, 2018.

[52] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[53] J. C. Alves dos Santos and E. L. Favero, "Practical use of a latent semantic analysis (lsa) model for automatic evaluation of written answers," *Journal of the Brazilian Computer Society*, vol. 21, no. 1, pp. 1–8, 2015.

[54] E. Del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "Gradeaid: a framework for automatic short answers grading in educational contexts—design, implementation and evaluation," *Knowledge and Information Systems*, pp. 1–40, 2023.

[55] J. Wang, H. Li, X. Du, J.-L. Hung, and S. Yang, "S-kmn: Integrating semantic features learning and knowledge mapping network for automatic quiz question annotation," *Journal of King Saud University-Computer and Information Sciences*, p. 101594, 2023.

[56] A. Hamza, N. En-Nahnahi, A. El Mahdaouy, and S. E. A. Ouatik, "Embedding arabic questions by feature-level fusion of word representations for questions classification: It is worth doing?" *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6583–6594, 2022.

[57] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," *Neurocomputing*, vol. 266, pp. 336–352, 2017.

[58] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.

[59] F. Bu, X. Zhu, Y. Hao, and X. Zhu, "Function-based question classification for general qa," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 1119–1128.

[60] M. Sarrouti, A. Lachkar, and S. E. A. Ouatik, "Biomedical question types classification using syntactic and rule based approach," in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, vol. 1. IEEE, 2015, pp. 265–272.

[61] K. Jayakodi, M. Bandara, and I. Perera, "An automatic classifier for exam questions in engineering: A process for bloom's taxonomy," in *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, 2015, pp. 195–202.

[62] S. F. Kusuma, D. Siahaan, and U. L. Yuhana, "Automatic indonesia's questions classification based on bloom's taxonomy using natural language processing a preliminary study," in *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE, 2015, pp. 1–6.

[63] M. Pota, A. Fuggi, M. Esposito, and G. De Pietro, "Extracting compact sets of features for question classification in cognitive systems: a comparative study," in *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*. IEEE, 2015, pp. 551–556.

[64] D. V. Paul and J. D. Pawar, "Use of common-word order syntactic similarity metric for evaluating syllabus coverage of a question paper," 2014.

[65] E. Sherkat and M. Farhoodi, "A hybrid approach for question classification in persian automatic question answering systems," in *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2014, pp. 279–284.

[66] L. A. Cutrone and M. Chang, "Automarking: automatic assessment of open questions," in *2010 10th IEEE International Conference on Advanced Learning Technologies*. IEEE, 2010, pp. 143–147.

[67] M. Razzaghnoori, H. Sajedi, and I. K. Jazani, "Question classification in persian using word vectors and frequencies," *Cognitive Systems Research*, vol. 47, pp. 16–27, 2018.

[68] J. M. Gomez-Perez, R. Denaux, D. Vila, and C. Badenes, "Hybrid techniques for knowledge-based nlp," 2017.

[69] J. Silva, L. Coheur, A. C. Mendes, and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artificial Intelligence Review*, vol. 35, pp. 137–154, 2011.

[70] S. Joshi, P. Shah, and S. Shah, "Automatic question paper generation, according to bloom's taxonomy, by generating questions from text using natural language processing," *International Research Journal of Engineering and Technology*.

[71] M. Pota, M. Esposito, and G. De Pietro, "A forward-selection algorithm for svm-based question classification in cognitive systems," in *Intelligent Interactive Multimedia Systems and Services 2016*. Springer, 2016, pp. 587–598.

[72] M. Zulqarnain, A. K. Z. Alsaedi, R. Ghazali, M. G. Ghouse, W. Sharif, and N. A. Husaini, "A comparative analysis on question classification task based on deep learning approaches," *PeerJ Computer Science*, vol. 7, p. e570, 2021.

[73] J. Liu, "Research on question classification methods in the medical field," *arXiv preprint arXiv:2202.00298*, 2022.

[74] F. Urrutia and R. Araya, "Automatically detecting incoherent written math answers of fourth-graders," *Systems*, vol. 11, no. 7, p. 353, 2023.

[75] G. Haisa and G. Altenbek, "Multi-task learning model for kazakh query understanding," *Sensors*, vol. 22, no. 24, p. 9810, 2022.

[76] H. Balla, M. L. Salvador, and S. J. Delany, "Arabic question classification using deep learning," in *Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System*, 2022, pp. 85–92.

[77] S. K. Gaddipati, D. Nair, and P. G. Plöger, "Comparative evaluation of pretrained transfer learning models on automatic short answer grading," *arXiv preprint arXiv:2009.01303*, 2020.

[78] S. Patil and K. P. Adhiya, "Evaluation of short answers using domain specific embedding and siamese stacked bilstm with contrastive loss." *Revue d'Intelligence Artificielle*, vol. 37, no. 3, 2023.

[79] M. Breja and S. K. Jain, "Analyzing linguistic features for answer re-ranking of why-questions," *Journal of Cases on Information Technology (JCIT)*, vol. 24, no. 3, pp. 1–16, 2022.

[80] A. Gašpar, A. Grubišić, and I. Šarić-Grgić, "Evaluation of a rule-based approach to automatic factual question generation using syntactic and semantic analysis," *Language resources and evaluation*, vol. 57, no. 4, pp. 1431–1461, 2023.

[81] C. Mallikarjuna and S. Sivanesan, "Question classification using limited labelled data," *Information Processing & Management*, vol. 59, no. 6, p. 103094, 2022.

[82] M. A. Sultan, S. Bethard, and T. Sumner, "Towards automatic identification of core concepts in educational resources," in *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 2014, pp. 379–388.

[83] K. Günel, R. Asliyan, M. Kurt, R. Polat, and T. Özis, "Dealing with learning concepts via support vector machines," in *Proceedings of the Seventh International Conference on Management Science and Engineering Management: Focused on Electrical and Information Technology Volume I*. Springer, 2014, pp. 61–71.

[84] G. Orellana, M. Orellana, V. Saquicela, F. Baculima, and N. Piedra, "A text mining methodology to discover syllabi similarities among higher education institutions," in *2018 International Conference on Information Systems and Computer Science (INCISCOS)*. IEEE, 2018, pp. 261–268.

[85] M. Badawy, M. A. Mahmood, A. Abd El-Aziz, and H. A. Hefny, "A text mining approach for automatic selection of academic course topics based on course specifications," in *2018 14th International Computer Engineering Conference (ICENCO)*. IEEE, 2018, pp. 162–167.

[86] M. Badawy, A. El-Aziz, H. Hefny *et al.*, "Exploring and measuring the key performance indicators in higher education institutions," *International Journal of Intelligent Computing and Information Sciences*, vol. 18, no. 1, pp. 37–47, 2018.

[87] S. Shaikh and S. M. Doudpotta, "Aspects based opinion mining for teacher and course evaluation," *Sukkur IBA Journal of Computing and Mathematical Sciences*, vol. 3, no. 1, pp. 34–43, 2019.

[88] O. Zahour, A. E. El Habib Benlahmar, and O. Hourrane, "Automatic classification of academic and vocational guidance questions using multiclass neural network," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 10, 2019.

[89] P. Pallegama, K. Kumari, D. Dissanayaka, A. Ravihansi, A. Karunasenna, and U. Samarakoon, "Evaluating teaching content and assessments based on learning outcomes," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, vol. 1. IEEE, 2020, pp. 299–304.

[90] H. A. Aalaa Abdulwahab and Y. H. Ali, "Documents classification based on deep learning," *Int. J. Sci. Technol. Res*, vol. 9, no. 02, 2020.

[91] A. S. Barb and N. Kilicay-Ergin, "Applications of natural language techniques to enhance curricular coherence," *Procedia Computer Science*, vol. 168, pp. 88–96, 2020.

[92] N. N. Vo, Q. T. Vu, N. H. Vu, T. A. Vu, B. D. Mach, and G. Xu, "Domain-specific nlp system to support learning path and curriculum design at tech universities," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100042, 2022.

[93] X. Li, A. Henriksson, M. Duneld, J. Nouri, and Y. Wu, "Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation," *Future Internet*, vol. 16, no. 1, p. 12, 2023.

[94] H. Fromm, T. Wambsganss, and M. Söllner, "Towards a taxonomy of text mining features." Association for Information Systems, 2019.

[95] Y. Liu, X. Yi, R. Chen, Z. Zhai, and J. Gu, "Feature extraction based on information gain and sequential pattern for english question classification," *IET Software*, vol. 12, no. 6, pp. 520–526, 2018.

Elia Ahidi Elisante Lukwaro is an assistant lecturer at The Open University of Tanzania and a PhD candidate in the School of Computational and Communication Science and Engineering at the Nelson Mandela African Institute of Science and Technology in Tanzania. He received a Master of Science in Information and Communication Technology at the Open University of Tanzania and the Bachelor's Degree with Honors in Computer Science and Software Engineering at Bedfordshire University in the United Kingdom.

Khamisi Kalegele is a senior lecturer at the Open University of Tanzania. He received his PhD in Computer and Mathematics Science at the University of Tohoku in Japan. His research areas of interest include promoting data using artificial intelligence and machine learning in the health, education, and governance sectors. Apart from being a senior lecturer, he is serving on the governing boards of the University of Dar es Salaam Computing Center, the Tanzania Forest Research Institute, and the Dar es Salaam Institute of Technology.

Devotha Nyambo is working as a senior lecturer and researcher at the Nelson Mandela Africa Institution of Science and Technology in Tanzania. She obtained her PhD in Information and Communication Science and Engineering at the Nelson Mandela African Institution of Science and Technology. Her research interests include agent-based modelling, agent-based simulation, machine learning, information systems (business informatics), information security, security