International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

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

Elia Ahidi Elisante Lukwaro[1], Khamisi Kalegele[2] and Devotha G. Nyambo[3]
[1,3]Nelson Mandela African Institution of Science and Technology,
Arusha Tanzania
[2]The Open University of Tanzania
Dar es Salaam, Tanzania.
E-mail address: eliaahidi@email.com, khamisi.kalegele@out.ac.tz, devotha.nyambo@nm-aist.ac.tz.

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

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 resources. The study utilized a methodical approach to survey the NLP techniques for extracting syntactic and semantic features for analyzing and understanding educational contents. The study identified the challenges and strengths of traditional and advanced approaches utilized in feature extraction. This review has benefits for stakeholders such as regulatory bodies, researchers, higher education institutes, and NLP researchers. The study provides NLP researchers with the current strengths and weaknesses of document analysis as well as the accumulated evidence skills for NLP-based application developers, which improves their ability to develop the appropriate algorithm and techniques for NLP tasks

Key Words: NLP; syntactic features; semantic feature; question classification; curriculum; educational content

1. INTRODUCTION

Recently, statistics have shown a significant increase in higher education institutes enrollment, while graduate unemployment is on the rise at both universities [1], [2], [3]. The advent of technology advances such as natural language processing (NLP) has piqued researchers' interest in developing a method to digitally evaluate educational processes that are linked to educational quality.

Higher Education Institutions (HEIs) carry out a variety of academic processes that not only ensure their competitive survival, but also determine the quality of service and education in general. These processes entail the generation of structured and non-structured academic content such as course materials, test questions, program details, and so on. Manual evaluation of quality aspects in such a context is difficult. Evaluating the quality of education is important for recognizing the effectiveness of education system for students' cognitive development and education's function in instilling citizenship ideas, value and attitude, responding to local and global challenges, creativity, emotional growth as well as analytical problem solving [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 quality 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: programs/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 presents the necessity of analytical automated methods. This is due to the fact that the aspects such as examination moderations, credit transfer, syllabi approval, or compliance can be a burden and incorrectly guided by human instinct if performed manually [8]. Approaches that have been used to evaluate the quality aspects of education can be categorized as human-based and automated. Total Quality Management (TQM) is one of the human-based approaches for managing quality that have been used in business as well as in education. it involves the set of principles and norms for improving services and product offered to the customer [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 is a subfield of Artificial Intelligence that is concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. When computers are given such abilities, a whole lot of quality assessment operations can be automated. Already, there are efforts to use techniques such as text mining and data mining, to check similarities of academic contents, audit educational information, evaluate standards of examinations questions, audit syllabus, evaluate factors that underlay

students' performance, visualize learning activities and many more [6], [10]; [11], [12], [13]. Relevant data and instruments are increasingly becoming digital, inviting applications of Natural Language Processing (NLP) techniques for effective assessment, control and evaluation the quality aspects in higher learning institutions.

This review focuses on investigation of the NLP techniques and associated challenges in extracting features that are used in analyzing and evaluating the quality of education based on syllabi and examination. The aim of this paper is to examine the state of the art of the applications of NLP in educational text analysis, as well as the strengths and limitations of the techniques applied. The remainder of the paper is organized in the following manner: Section 2; Literature review, Section 3; Proposed approach, Section 4; Results and Discussion, Section 5; Draws the conclusion.

2. LITERATURE IN REVIEW

The application of analytical automated technology to educational data is gaining popularity. Several reviews that looked at various aspects of education provide the insightful evidence on the issues regarding NLP's techniques, that are important for evaluating the educational data. These reviews have used various approach that provide the empirical evidence to the body of knowledge. The comprehensive review approaches include the review by [14], whereby an integrated approach to feature extraction, such as keyword, headword, syntactic, and semantic extraction, was presented to classify questions contains keywords assigned to more than one level of Bloom's taxonomy (BT). [14] investigated statistical approaches for question answering systems, information retrieval, and educational environments using machine learning approaches such as support vector machine (SVM) and other classifiers. The study acknowledges the importance of semantic and syntactic extraction in archiving reasonable accuracy with SVM classifiers in classifying questions in question answering systems and information retrieval, but finds less performance in educational settings. According to [15] study, NLP features like lexical and semantic matching with machine learning techniques like SVM increase classification accuracy in question answering systems. However, machine learning baseline is significantly impacted by the quality of the dataset's domain, indicating the necessity for more research on machine learning's cross-domain application.

The systematic review approaches include reference by [16], who conducted a study of automatic question classification methods based on computer programming exams. [17] reviewed NLP techniques and suggested some techniques, including using lemma instead of words, to improve the Unified Medical Language System (UMLS). The work also suggests that medical students' medical documentations be improved by utilizing a spell-checker that is enhanced by NLP that offers real-time educational input. The review by [18] analyzed techniques and algorithms for question classifications. The study indicates that the SVM is the major machine learning technique utilized in classification, while the main techniques for feature extraction and selection are Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). The BOW technique was used in response processing, and it was verified that SVM was one of the best algorithms for this type of problem. The systematic review by [19], whereby the articles between 2015 and 2019 regarding auto question generation are reviewed, the conclusions note the need for additional experimental reporting using standardized metrics and greater research and evaluation of plainly evident approaches. The study by [20], which examined a variety of techniques for educational text mining, indicated that NLP was the most effective text mining tool for the education industry. They do, however, reveal that the majority of reviewed articles place a greater emphasis on the outcome rather than the process. As a result, systems are created that are accurate but lack interpretation.

There has been a significant amount of research on NLP, text mining, and machine learning techniques as a result of the increasing volume of data produced by educational processes as well as the desire for competence and quality. Most studies have utilized systematic as well as comprehensive methods-based approaches to study the various educational issues related to NLP techniques. The well-formulated methodic approach was used in this study to review the NLP techniques as well as their associated strengths and challenges in extracting educational data, specifically in curriculum and examinations. The contribution of this work to the body of knowledge includes the analysis of a wider range of innovative publications to learn the state-of-the-art of NLP in processing education data. An in-depth discussion of the strengths and challenges of the techniques used in educational data processing, due to the fact that NLP applies to a wide range of fields and various techniques are used in many other fields, this article may be beneficial in fields other than education. Based on our extensive literature review, we accumulate the body of evidence that is useful for providing NLP researchers with the knowledge they need to adopt the best algorithms and techniques for NLP tasks while also educating NLP-based application developers on the most recent strengths and challenges of document analysis.

3. PROPOSED APPROACH

A. *Data Sources*

The review examines NLP techniques as well as related strengths and challenges in processing educational data such as examinations, curriculum, and educational content. The following scientific repositories were used for data collection: Science Direct (https://www.sciencedirect.com/), and Google Scholar (https://scholar.google.com/). The repositories include several large-scale studies from different journals. The underneath section described in details algorithm used for data retrieval from these sources.

B. *Search Query Strategies*

Search queries are created by combining keywords by using Boolean operators. The search query as indicated in Figure 1 is generated from three lines of search terms, which represent NLP techniques and the types of documents that

should be included in retrieval articles. The first line of keywords (K's 1) is included in search queries based on NLP syntactic structure features (SF_n); lexical and syntactic analysis features, which represent syntactical features such as sentence-splitting, morphological analysis, tokenization, phrase structure, stemming, POS, and others dealing with the relationship between syntax and grammar. The second line of keywords (K's 2) represents semantic features (SM_n), which deals with the meaning of the words and their context relations within a sentence. Furthermore, the second line has features that perform sophisticated semantic representation of text data for text analysis, i.e., topic generation and document classification. The third line of key words (K's 3) represents the type of document (DT_n), such as educational content, questions, or curriculum. The comprehensive list of search queries was developed using a series of NLP techniques: from syntactic structure, semantic representation, and advanced document analysis; which generates the results (R_{1-n}) of articles from several repositories. This search query was the most effective at accessing the chosen databases and producing results that were pertinent after testing several rearrangements with numerous Boolean configurations. Although the aforementioned sets of NLP techniques may overlap with each other, the main focus was on identifying the core techniques that are useful for content understanding and analysis in a given document, such as extracting relevant features from unstructured raw text data and converting them to a more structured form of representation for machine learning.

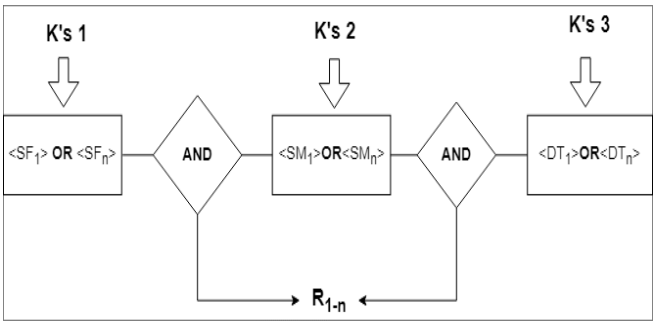

Figure 1. Flow diagram represent the combination of keywords in search queries

C. *Procedures used in data retrieval*

Based on the aforementioned query generation, Table 1 shows the search query terms used for Science Direct and Google Scholar Repositories. The search query terms are formulated by following the combinations of search terms illustrated in the search query strategies section.

Table 1. 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.

Following execution in both repositories, a sizable amount of data was produced. For the Direct Science repository, the articles were refined before initial screening. The refined by the "articles type" tool, which was utilized to select the "research articles" options, led to discarding irrelevant articles such as reviewed articles, encyclopedias, and etc. Furthermore, the results were refined through the "the subject area" option, which in our case, a computer science subject was selected. Moreover, the "publication title" option was used to select the first relevant group of journals. A total of 1,630 articles were retrieved from the following Direct Science repository journals: 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. For the Google Scholar search platform, a total of 1,180 articles were retrieved, whereby 59 articles were left out as they were review articles and remain with 1,121 articles. Furthermore, 9 articles were selected from relevant reference lists.

D. *Inclusion Criteria*

The retrieval algorithm led us to remain with a total of 2,760 articles from both repositories for further screening. These were then screened using inclusion criteria listed in Table 2, which include details of publications such as range of years, title, and language. Furthermore, the abstract and subsequent contents were scanned, whereby articles with relevant keywords as well as content that specifically included NLP techniques in evaluating educational content such as examinations, syllabi, and curriculum were included.

Table 2. Inclusion Criteria

SN	Factor	Inclusion Criteria
1	Year	2010 ~ 2021
2	Language	English
3	Types of Publications	Peer-reviewed, working papers and books
4	Title	Relevant concept per study
5	Abstract	Keywords related to study
6	Text Screening	NLP techniques in evaluating educational contents such as examination and curriculum.

Consequently, based on inclusion criteria, 2,529 articles were initially excluded. The second phase included screening the keywords and abstracts, whereby 156 were excluded. Furthermore, the full text screening in the third phase resulted in the removal of 15 articles which contained unrelated educational data based on examinations, curriculum, and educational contents. As shown in Figure 2, the 60 articles were chosen to be included in our study as they matched the metrics.

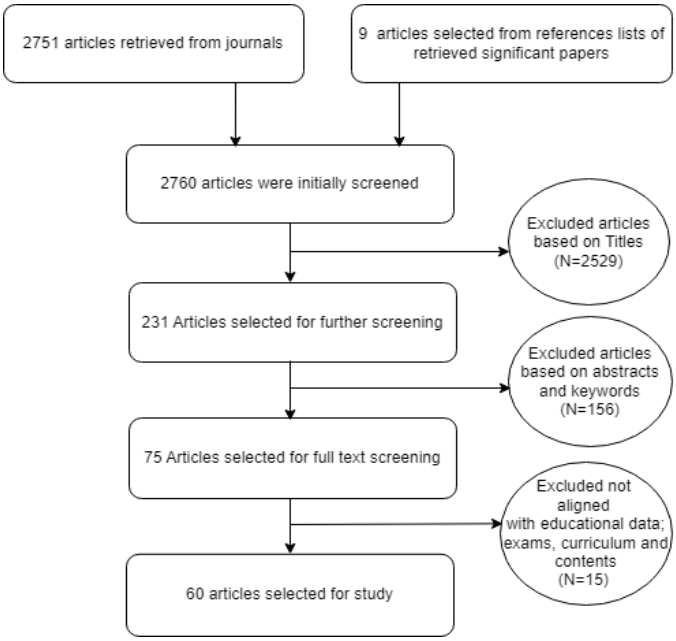

Figure 2. Inclusion and Exclusion Criteria

A. *General analysis of the selected articles*

The remaining articles (60) that were relevant to this study were reviewed in search of specific cases of NLP applications. The cases were thoroughly analyzed to identify their similarities in order to simplify their presentation. The details of the general analysis of the retrieved articles during the years 2010–2021 are depicted in Figure 3. The articles are categorized as those with techniques that evaluate exam questions and various assessments according to the BT model. Others include techniques that evaluate exam questions and other assessments per factors such as answer categories, pattern matching, and syllabus coverage. Other articles include studies on curriculum and syllabus evaluation techniques, as well as other educational documents such as lecture contents. These articles describe techniques for evaluating educational documents to determine their quality according to various standards. More studies have been done on evaluating examinations according to BT and other quality factors than curriculum and other educationally related data. This could be due to the challenges of evaluating examination questions, as mentioned by [21], or the importance of high-quality examinations as the primary means of assessing gained skills or learning outcomes. There has been very limited research on curriculum evaluation.

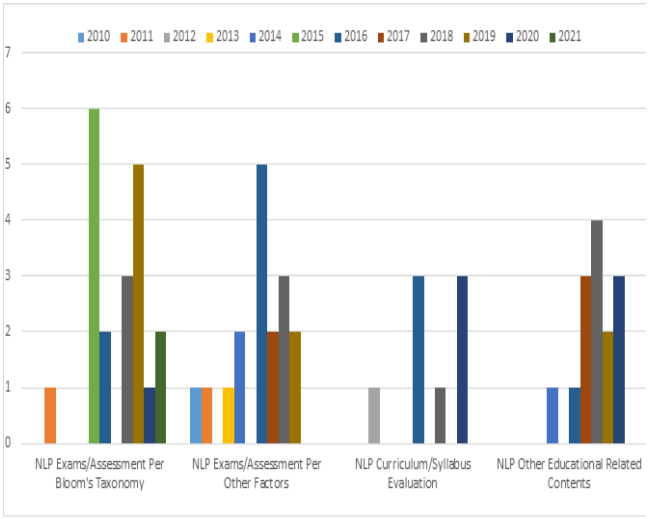

Figure 3. Analysis of the retrieved articles from 2010 to 2021

4. RESULT AND DISCUSSION

There are common similarities in the phases that involve the processing of text in NLP application by these articles (Fig. 4). These phases include techniques for converting, representing, dimension reduction, feature selection, feature extraction, and classification, all of which are aimed at processing natural language, i.e., text, into a format that can be further processed by machine learning. The techniques vary in phases as well as in their application to the hands. However, most studies choose their applicability based on aspects such as completion of tasks with minimal processing costs, semantically and syntactically portrayed performance, and efficiency.

NLP with text mining and machine learning is a preferred technology in evaluating aspects of higher education quality. This is because they include features that aid in the comprehension of textual content, as well as implementing techniques based on natural language that provide an interpretive interface between human and machine [22]. Also, it is enriched with a plethora of toolkits that enable the creation of powerful application without the need to start from scratch [23]. Traditional or count-based features, use mathematical and statistical methods such as the BOW, TF-IDF, N-grams, and topic modeling, which use models such as SVM, Random Forests (RF), Decision Tree (J48), and Naive Bayes (NB) for document analysis. Other categories include deep learning approaches, with models such as Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM), and Gated Recurrent Unit (GRU), which use Continuous Bag of Words (CBOW), Word2vec, skip gram, and other techniques [24]. Even though deep learning seems to work better than traditional methods, several hybrids have been made to improve the old methods and increase their effectiveness [25].

A. *NLP techniques for feature engineering based on text input data analysis*

The literatures have addressed the applications based on text mining and NLP on educational data analysis. The applications utilized machine learning algorithms, to intelligently evaluate the quality of academic data such as examinations and curriculum [12]. NLP techniques are stated to address text mining drawbacks in preprocessing text data and extracting relevant features that demonstrate a greater comprehension of text contents, which in turn improve text mining algorithms and deliver the best results for the task at hand [26]. The document analysis task, which saves as the fundamental for many activities such as extracting important insight from text and application in numerous fields, is one example of how NLP may be used for automated text analysis.

The document analysis task, which saves as the fundamental for activities such as extracting important insight from text and application 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. The text preprocessing process includes techniques such as tokenization, lower casing stemming, etc., which clean and transform text documents for upfront processes. The text representation techniques such as BOW, N-gram, and others convert text into a mathematical computational format, sometimes known as "feature extraction," which is important for classification processes. The classification process such as SVM, NB and others, classifies the represented text for further tasks or applications. This study examines the issues of technique and feature extraction from educational data, with an emphasis on exams, content, and curriculum. Document analysis plays a significant part in accomplishing the task, and it consists of three processes: preprocessing, document representation, and classification.

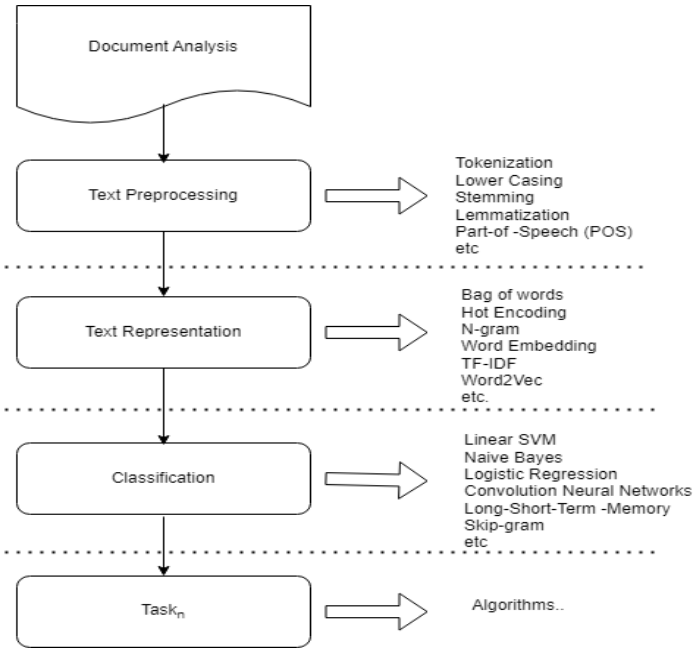

Figure 4. Text Analysis Processes and Techniques

The tasks that NLP can accomplish include matching, classification, translation, structure predication, and the sequential decision process [27]. These tasks require techniques for extracting features based on syntactic and/or semantic features and placed them in a format that can be used by the task at hand [28] i.e., matching or benchmarking. [22] describes and categorizes 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-phrased-based features and parse-tree-based features. These reveal significant information regarding sentences or phrases' meaning
[29], entity relation features which main tasks is to find and characterize semantic relationships between text entities [30]. On the second line are pure statistical features, which include patterns such as statistical phrases, frequent word sets, and frequent sequential word patterns. These extract patterns consist of a large number of words from text input that can be utilized as a feature [26]. Third line of categories includes; singular value decomposition (SVD), probabilistic topic modeling, probabilistic latent semantic index, LDA. These include features that conduct dimension reduction to reflect concepts rather than raw terms, as well as a probabilistic model to find phrase co-occurrence patterns in a collection of documents that correlate to semantic subjects [31].

Furthermore, [25] describes NLP techniques for feature engineering by categorizing them into traditional and advanced feature engineering models. Traditional or count-based feature engineering extracts features from text using mathematical and statistical methods such as the BOW model, TF-IDF, N-grams, and Topic Modeling. The aforementioned features can be used to evaluate the similarities of documents and other domains, including search engines, document clustering, and information retrieval. The traditional methods lose additional information such as semantics, structure, and context of words in documents. Nevertheless, advanced feature engineering leveraged the weaknesses of traditional based features to develop vector representations of words. Advanced feature engineering uses predictive techniques, such as neural network language models that analyze word sequences and forecast words based on their neighboring words, also known as embedding features [25].

B. *Overview of NLP techniques for examination analysis*

Examinations are the most common method for evaluating students' cognitive capacity in universities. Several studies have been undertaken, to automate the task of evaluating the quality of examinations. Some studies present the use of BT [32], [33], [34], [35], Solo [36], construct question-answering systems [37], and match the examination with learning objectives [38], are just a few of them. The classification process is critical for completing the aforementioned tasks;

the underneath sections offer an overview of the techniques used in examinations classification.

1) *Techniques for Preprocessing Examinations*

Reviews studied tend to concentrate on aspects that clean and transform data in the format that can be utilized by upfront processes [35]. These include the techniques that transform examination questions by removing diagrams and symbols, descriptive text prior to the questions and words with less than three letters, punctuation removal, and remove non-Unicode characters [38]. Other techniques include changing characters to lowercase, removing punctuation marks, numbers, tokenization which breaking content down into manageable chunks or tokens, punctuation removal, phrase segmentation, stop word removal, part-of-speech (POS) tagging and parsing [39], [33], [36], [34], [40], [12], [41], [42], [38], [43]. The majority of the techniques discussed above deal with the appearance of words by defining each word-form using lexical and contextual information, while semantic and pragmatic issues are addressed later in the processing stage. Another technique is normalization, which eliminates word form variability to a similar style [12], [41], [44]. Some studies used stemming, such as [43], [41], [42], [36], [12], and [44], but others did not. However, lemmatization is preferred in the case of question classification, according to [44], because it uses the WordNet lemmatizer to find legitimate root words which is important in extracting semantic information from exam questions, because the meaning of words is important in extracting semantic information from examination questions [45]. Furthermore, [40] use techniques such as tokenizer, which also changes lowercase to uppercase, breaks sentences into words, and returns frequently used phrases while omitting less frequently used ones. Additionally, label encoding is used to turn labels into machine-readable forms, while pad sequence is used to ensure the maximum sequences in a list have the same length. Preliminary processing is a crucial step in preparing text input for subsequent machine learning processes. Several issues may be traced back to the preprocessing steps, including the removal of crucial symbols like U for voltage and I for current, in order to address the new custom symbols introduced to prevent them from being eliminated as stop words [42]. The process of lemmatization, which entails translating a word to its original roots, may be difficult for low-resource languages like Swahili to construct a structured semantic relationship between words, resulting in lower performance by misrepresenting the correct meaning of the word in question [46], [47], for high resource languages, such as English, WordNet lemmatizer is used to retrieve real root words marches [36]. It may be difficult to construct a structured semantic relationship between words in low-resource languages such as Swahili, resulting in lower performance by misrepresenting the correct meaning of the word in question [46], [47]. For high-resource languages, such as English, the WordNet lemmatizer is used to retrieve real root words via marches [36].

2) *Techniques for Text Representation*

Most studies don't make much of a distinction between preprocessing and text representation. Text representation encompasses techniques for converting text into a numeric vector that may be evaluated semantically or syntactically [48], as well as dimensionality reduction via feature extraction, which creates additional features that retains the useful information., or feature selection, which keeps a subset of the original features [45]. Feature selection and extraction have been utilized in a number of studies as significant ways for representing text in a format that can be utilized in upfront process i.e., text classification.

Feature extraction includes the technique such as; unigram, wh-words, word shapes, tagged unigram, head words, related word group, hypernyms, tagged unigram and bigram [37], [49], [50]. Other techniques include BOW (Pintar *et al*., 2018), verbs or keywords extraction (Dhainje *et al*., 2018; Jayakodi *et al*., 2016), [34] weightage using TF-IDF and then LDA generate unique topic for each question based on stemmed words [38]. The N-gram, or unigram represents words, while the Term Reference is used to count the amount of verbs and nouns [39], [44]. Other techniques include enhanced E-TFIDF [41], modified TF-IDF to TFPOS-IDF [12] (Mohammedid, M., & Omar, 2020), grammatical patterns that relate to the text's words [51], bag-of-concepts [48]. N-gram, or unigram represents words, whereas the Term Reference is utilized to count the number of verbs and nouns [39], word2vec embedding vector which include the variant of text representation such as continuous bag of words (CBOW) that generates word representations by identifying a center word from a window of selected context words and skip-gram which constructs word vector representation by identifying the context words around a given word [52]. Feature selection include latent semantic analysis which uses singular value decomposition, a mathematical approach, to search unstructured material for hidden links between phrases and concepts [53], mutual information which choose the most key features from the original data collection, chi-squire statistic that choose features that are strongly reliant on the response. and odd ratio which select selecting on appropriate feedback words [32].

The findings show that there are strength and significant limitations to the text representation techniques. Table 6 represents same of the salient features including strength and limitations of NLP techniques. According to the findings, common and widely used techniques such as bag-of-words have drawbacks such as ignores the semantic, conceptual, and contextual information in the text, as well as having high dimensionality and sparsity issues [48], also, fail to preserve the necessary proximity information as the number of unique words grows [54]. In research like [38], the words in questions are represented by assigning weightage using the TF-IDF approach, the limitation arise when different words are given the same weighting and the experiment does not cover all Blooms Taxonomy levels. Several factors have been presented to determine the performance of techniques, including data size, which states that large data sizes maximize performance while small data sizes minimize performance, and dimensionality, which states that low dimension improves performance by reducing computational

cost and storage space, and thus improves algorithm performance [53].

1) *Approaches based for 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 3 is a summary of some of the most common ways to classify and rate exams, as mentioned by different authors.

Table 3. Questions classification summary based on various criteria and techniques

Authors	Criteria	Evaluate attributes	Classification techniques
[55]	pattern matching it include six categories: fact, list, reason, solution, definition, and navigation.	Rather than being classified according to their contents, questions are classified according to their functions.	Multi-Layer Neural Network (MLN)
[56]	Classify Biomedical Question into 3 categories; Yes/No, Factoid and Summary Questions.	How, which and what types questions categorized as factoid and summary. Where is categorized as factoid. Yes/No categorized as Yes or No question	Using a question pattern, an algorithm was created to classify questions into predefined categories.
[51]	Classifying syntactically into six different categories, which are: causal, choice, confirmation, factoid, hypothetical and list.	Causal (explain events), choice (Questions mostly provides choices), confirmation (Yes-No Questions), factoid (facts, current events, opinions, and recommendations), hypothetical (general understanding of a situation) and list (List of facts or entities).	Support Vector Machine (SVM), Random forests (RF), Decision Tree (J48), Naïve Bayes (NB)
[38]	Questions Classification based on objectives or learning outcomes	Automatically label practice opportunities based on the course creators' anticipated learning outcomes.	Support Vector Machine and Extreme Learning Machine
[40], [35]	Questions Classification based on Bloom's Taxonomy (BT)	Questions categorized per BT taxonomy cognitive levels such as Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation.	SVM, NB, Logistic Regression, and Decision Trees, rule based, RNN, LSTM
[57]	Questions are categorizing per syllabus coverage	Examine a question paper's coverage of the syllabus.	Performed matrix similarity via matrix representation vector an extension of VSM
[58]	Question classification approach for closed-domain question answering systems	Determine the overall performance of Course and Fine grained for examinations answering system	Combined rule based and machine learning approaches

The studies classified question classification into ruled-based, machine learning, hybrid, statistical, and deep learning approaches, which use a variety of feature extraction and selection methods to classify the questions based on expected metrics. Table 4 indicates the categories of classification as well as strength and limitations.

Table 4: Classification approaches and challenges faced

Authors	Classification Approach	strength	Limitations
[35], [33]	Rule-based approach	Quite accurate, despite the fact that they are time intensive and demand hard human efforts.	Time consuming, tedious, poor performance when compared with ML, and when using BT to classify questions. Effective for categorizing questions into knowledge levels.
[45], [58]	Machine Learning-based Approach	Using ML, a high-performance question categorization system can be developed automatically, utilizing hundreds of features.	The number of dimensions in an SVM model causes a lack of clarity in the outcomes as well as computational complexity, whereas in a Nave Bayes model, accuracy is reliant on the extent of training data.
[59]	Statistical-based Approach	Able to work with large amounts of data	Relied on conceptual understanding rather than semantic evaluation.
[58], [60]	Hybrid-based Approach	Handle the inherent limits of each particular technique while simultaneously utilizing the benefits of each	The classification of Persian inquiries, as well as failing to continue with the process after encountering an English term in a Persian language inquiry, necessitates the use of human resources and can be challenging.
[40], [27]	Deep learning Approach	It supports the automatic learning of multi-level	The model requires a large amount of data, making it

		feature representations and has high precision when trained on huge data.	unsuitable for exam data. It's also expensive to train, due to high computing costs and a lack of theoretical grounding, making selecting the right deep learning tools difficult.

Previous techniques depended on a statistical and probabilistic approach, which in turn relied on conceptual understanding of text, but with the advent of NLP, the previous focus was switched to semantically evaluation, allowing for the natural evaluation likeness of language to be captured [59]. Statistical approaches include linear methods such as SVM and probabilistic topic models, as well as non-linear neural networks, both have strength and limitation, but there is an effort to combine them in order to have the best performance [61].

Other studies focus on evaluating the performance of various method these include: [37], [49] and [50]. The studies evaluate the classification's performance by combining features exctracted from lexical, syntactic, and semantic. The authors in [49] combined the followings; lexical features i.e. unigram (U) and word shape (WS); syntactic features i.e. headword (HW) and question category (QC), and semantic feature; hypernyms (HY). The study by [50] used lexical features such as unigram, Bigram (B), and word shape; syntactic features such as head word; and semantic features such as query expansion (QE), question category, and related word (RW) to design the feature known as question patterns (QP), which he combined with the other mentioned features into a unique form. Table 5 shows that coarse grain yields better outcomes by 1% for [49] and 0.5% less for fine grain when compare to [50]. However, coarse grain yields less than 3% for [37]. According to [49] study, the hypernym and quation category features enable to uncover relationships in naturally occurring text, allowing it to perform better in NLP tasks such as categorization.

Table 5: Compare the study of features, algorithms and performance

Study	**Feature 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%

The majority of early approaches to questions classification depended on rule-based procedures, with classification rules created manually [62]. [62] use a rule-based approach to find the performance of features, rule-based questions classifier manipulates and generates features that can be used with other features within SVM to improve performance. The results show that while headwords features (H) on their own produce poor results, but when combined with category features (C) for course granularity, they produce better results, and when combined with unigrams (U), the classifier delivers the best results [62].

Other studies on questions classification are main focus on classifying questions per Bloom's Taxonomy. Various techniques have been reported including apply preprocessing operations with the use of a rule-based strategy for categorizing queries, with a weighted category for all overlapping keywords [35], however, owing to the variability in background knowledge of each domain, this technique may result in inconsistencies, resulting in poor classification performance. [41] modified TF-IDF by enhanced E-TFIDF whereby impact factor introduced higher calculation to the series of verbs, nouns and adverb over others, then the results were analyzed using SVM, NB, and KNN classifiers. The results showed that the enhanced E-TFIDF produces better results than the others [41]. Another study compares the performance of three features: TF-IDF, TFPOS-IDF, and W2VTFPOS-IDF. The results show that by modifying traditional TF-IDF to TFPOS-IDF, it focuses on giving verbs higher priority over other words, and that with W2VTFPOS-IDF, it provides the context of questions as well as high-quality feature vectors representation. The average recorded results with different classifiers were; Logistic Regression and Support Vector Machine were 71.1 %, 82.3 %, and 83.7 %, respectively, whereas the records for 600 questions in the same classifiers were 85.4 %, 89.4 %, and 89.7 %. The authors in [44] combined syntactic features like part of speech tagging (POS) with semantic features like WordNet and the Lest algorithm to classify examination questions into Bloom's Taxonomy. The N-gram, or unigram, used to represent words, whereby Term Reference is utilized to count the number of verbs and nouns, classifiers SVM, NB, and J48, were used. The results revealed that classifiers with combinations of features outperformed those without, and scored higher on the f-measure, with SVM coming out on top. The authors in [63] used techniques including verbs recovered from sentences and stemmed by Lancaster stemmer, whereby the POS tagging was used to generate the sentence skeleton and WordNet to identify the correct root of word, resulting in a 72.9 % accuracy. The study by [53] used 1,250 questions from programming and other courses to automatically classify the questions per BT. The model uses latent semantic analysis with SVM after preprocessing and gets a score of 86%, while preprocessing with SVM but not LSA gets a score of 96.65%.

The authors in [39] use TF-IDF and NB to classify examinations based on BT cognitive levels. Preprocessing techniques such as dataset labeling, tokenization, stemming, and filtering are used, as well as feature extraction utilizing the TF-IDF technique on a series of words, characters, and N-grams. With the TF-IDF, N-gram approach obtained the highest accuracy precision of 85 %. The dataset used for this study included mid-term and final exams from Telkon University's Department of Information Systems. Furthermore, several studies classify question per answering system. The study by [64] used SVM as a classifier and a set of low-dimensional lexical and syntactic features for

summarizing the content of a larger set and questions classification. The obtained accuracy was 89.2% for course classes and 82.4% for fine classes, which is less compared to result reported by [49] and [50]. This could be related to the training data set, but it is worth noting that semantic features were left out of the [64] study. The authors in [51] classified questions using grammatical structure, syntactic features, and other techniques, as shown in Table 4. The J48 decision tree classifier outperformed the other classifiers by 91.1%. The authors in [65] applied four different deep learning approaches; CNN, GRU, LSTM, CNN-GRU, and CNN-GRU for question classification. Word2Vec embedding vectors, such as Skip-gram and CBOW, were employed in the 5,400 questions for training and 600 questions for testing. Word2Vec quickly learns the semantic and syntactic links between words in a document, improving the performance of classification models. In skip-gram mode, CNN-LSTM and CNN-GRU approaches outperform CBOW by 93.7% when the maximum of 300 characteristics is applied. When the English and Turkish data sets were compared, the English data set got 94.4% accuracy using the LSTM technique, but the Turkish data set didn't because it was too complicated.

C. *Overview of the educational document analysis and incorporate tasks*

In this review, educational documents are used to represent course material, syllabi, and curriculum. Document analysis techniques are similar to question classification; however, the research reveals that questions classification faces same challenges compared to document classification due to the shorter length of questions. Thus, the two studies should be conducted differently [21]. The NLP and text mining research on educational document analysis includes studies that identify learning concepts from learning resources [66], [67]. The study by [48] introduced bag-of-concepts to address the traditional BoW in document classification tasks, notably in the text representation process, i.e., increased dimensionality and sparsity concerns. While others evaluate the similarity of syllabi among higher education institutions using the UNESCO knowledge area classification [68]. The authors [69] automatically relating the topic with the course book and checking for missing parts in course specifications, while [70] identifying the course's Knowledge Performance Indicators. The study by [71] deploying 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 [72] evaluate teaching material and assessment based on learning outcome, while [73] evaluate the coherence of an academic curriculum. Furthermore, the study by [74] implement multi-sentence classification on a large number of documents using CNN. According to the studies, NLP-based educational document analysis covers a wide range of tasks, when compared with examination analysis studies whereby most researches are centered on question classification using the Bloom taxonomy, as well as classifying per examination answered systems.

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

Text preprocessing, representation, categorization, and finally completing the required task are all linked to various techniques and approaches (Fig. 4). In [67], the study deploys preprocessing techniques including removing mathematical formulation symbols, variables, and numbers in context; parsing sentences and paragraphs; removing punctuation marks and special characters; changing characters to lowercase; and excluding white space characters; then using the document's n-gram and TF-IDF as extracted features in order to train the SVM in the determination of whether or not the document contains a learning concept. Furthermore, dimension reduction techniques such as singular value decomposition Principal Component Analysis (PCA), and Multi-Dimensional Scaling (MDS) were utilized to eliminate feature space noise. The study found that when dimension reduction strategies are utilized, the system's accuracy is low, meaning that most learning ideas are missed. As a result, SVM is relatively limited when data is unbalanced. Another study [66] determined the core concept from educational resources by examining how closely it is related to the domain topic. The authors in [68] used SVM classifiers, preprocessing techniques, and TF/IDF and LSA used for dimensionality reduction as well as TF/IDF for feature selection. Then, latent semantic analysis is used to cover the relevant features, and cosine similarity methods are used to evaluate the similarity of 1,442 syllabi of computer science courses from Ecuadorian Higher Education Institutes, whereby the classification is based on UNESCO knowledge areas. However, the study by [60] encountered challenges including a high degree of similarity in courses with different contents but with the same topics. The study by [69] deploys text mining and NLP to automatically relate the topic with the course book and check for missing parts in the course specifications. The techniques used are tokenization, stop word removal, and case transformation in the preprocessing phase; N-gram for keyword extraction or targeted words; and term frequency and N-gram to select and analyze the contents of the course topic versus the course specification. The authors in [70] use text mining techniques to identify the course's Knowledge Performance Indicators (KPIs) using preprocessing techniques such as tokenization, stop words removal, and stemming, as well as extracting synonyms of words and keywords that depict Intended Learning Outcomes (ILOs), and then calculating term frequency based on knowledge and understanding, intellectual skills, professional skills, and practical skills. The authors in [71] evaluate teachers and courses, the supervised ML extract general topic while NLP techniques were utilized to locate specific and relevant opinion terms for which feedback was given. The used techniques include Apache OpenNLP for preprocessing, TF-IDF was used in String2WordVector, a feature extraction tool, Java's standard core NLP API extract the required feature. The Naive Bayes Multinomial classification was deployed for text classification. The processes achieved a recall and precision of 83% and 84%, respectively; the limitations included the system's failure to resolve new input words, wrong English words, as well as the

chance of assigning polarity to words that do not exist in SentiWordNet.

The study by [75] used the Multiclass Neural Network approach to classify the category of academic and professional counseling queries according to Holland's RIASEC topology. RIASEC represents six personalities, Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C) that correspond to Career Choice or educational program environment. Multiclass Neural Network outperforms prior approaches such as SVM and feature-based classification algorithms by allowing it to process datasets with complex issues, such as those in the biomedical field. Preprocessing and feature engineering models were utilized, in which raw text was converted to integers and the integers were used as the model input. The dataset was split into 70% for training and 30% for testing, and then the Multiclass Neural Network for

classification was achieved. Results indicated that Multiclass Neural Networks perform better than other algorithms of ML. The authors in [72] employed four components, each of which assesses teaching material and assessment based on learning outcomes. The components are: analyzing learning outcomes and levels in Revised Bloom's Taxonomy; evaluating lecture material fairness; analyzing question paper fairness; and evaluating practical session fairness. The data set includes 600 learning results from several modules that were preprocessed with tokenization, stop word removal, text conversion to lowercase, and the Natural Language Toolkit (NLTK). The Recurrent Neural Network with LSTM architecture was utilized to classify module lessons, learning outcomes, summarize lecture power point presentation (ppt), and check slides that covered learning outcomes. For multiple text classification on a large document, [74] examined three models: CNN, standard LDA, and modified LDA with TF-IDF. According to the findings, the improved LDA improves accuracy from 60% to 74.44% and reduces time from 4.04 to 3.02. Despite the fact that both two LDAs had better time management, CNN outperformed the other two modules by 94.7%. Furthermore, the preprocessing techniques used include tokenization, deleting words with less than three characters, removing stop words, lemmatizing, and stemming. [24] proposes the text similarity by considering semantic sequence of words rather than syntax. The preprocessing phase include tokenization, lowercasing, and stemming of short text. The approach considers the word level coherence by hybrid method of dependency parser and lexicon embedding that linked to the external resources such as ConceptNet. Then the sentence pairs' similarities were calculated using the bag-of-words (BoW) vector. To identify the entities and concepts in the document, [48] use techniques such sentence segmentation, word tokenization, and POS tagging, while entity sense disambiguation is employed to deal with polysemous terms and improve text representation. Bag-of-concepts were developed to address the problem of traditional BoW in document classification tasks, particularly in the text representation phase, where increased dimensionality and sparsity concerns were raised. 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.

Table 6. Critical Salient Features Introduced in Articles

Year &Authors	Topic Covered	Data Used	Used Techniques	Strength of the techniques	Limitation of the techniques
[37]	Examination Questions Classification by weighted combination of lexical, syntactic and semantic features	A total of 6000 questions from the University of Illinois Urbana Champaign (UIUC) were mapped to the questions taxonomy	Unigram, wh-words, related word group, word shapes, tagged unigram, bigram, head words and their hypernyms as features on WordNet. Using SVM classifier	The techniques have a lower computational cost compared to the state of the art at the time	The questions were categorized according to a question taxonomy. The challenge with linguistic analysis levels when creating a feature taxonomy is that there isn't always a one-to-one relationship between a feature and an analysis level [76]
[35]	Automated analysis of Exams based on Blooms Taxonomy	Programming 70 (training set) examination Questions. Test data 30 Questions	Preprocessing; stop words removal, stemming, lemmatization and POS tagging. NLTK tagger for text representation. A	Categorized the questions per cognitive level	For keywords that are overlapping, the weight category used. This may lead to inconsistency due to variety of knowledge levels in Bloom's taxonomy

		used in final exams	rule-based approach is utilized to discover important keywords and verbs that determine question's categorization..		
[49]	Question Classification using Semantic, Syntactic and Lexical features (classification based on concise answers)	Different training set of questions range from 1000 to 5500	Preprocessing; stemming and stop word removal. Text representation; bag-of-words, unigram, Headword, Hypernyms, Word shapes, Questions Category. Using Linear SVM classifier	Probe and identify the questions to the probable category hence increase the performance	syntax feature extraction typically comes with a heavy computational cost [77], hence, time consuming, as it utilized much resources due to the parser process. Also Bag-of-words technique ignore the semantic, conceptual, and contextual information in in the text, as well as having high dimensionality and sparsity issues [48], also, fail to preserve the necessary proximity information as the number of unique words grows [54]
[64]	Classification of questions and large information using SVMs, forward-selection algorithm, based on new introduced features	UIUC benchmark dataset, consists 5452 training questions and 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 feature hence reduce dimension,	It has significant shortcomings, such as poor discrimination in the case of some Principal-Wh-Words and some classifiers' reliance on the training dataset.
[43]	Classifications of Exam Questions Using Linguistically-Motivated Features: A Case Study Based on Bloom's Taxonomy	Questions from Najran University, the computer science program 600 questions	Preprocessing; tokenization, POS, Stemming, Lemmating, N-gram. Text representation;Unigrams, Bigrams, Trigrams, POS Bigrams, POS Trigrams, Word/POS Pairs, and Stem N-grams. Number of classifiers were trained and tested; NB, Logistic Regression, SVM, Decision Trees	Combination of all features produce better performance with SVM and Logistic regression, in addition more N-grams such as bigrams and trigram substantially improve performance by preserving local word sequence ordering	The linguistic technique does not establish a link between a feature taxonomy and a certain analysis level. The study as well did not take into account the semantic structure of examination questions, resulting in lower accuracy or bit improvement
[66]	identifying core concept from educational resources	340 sentences from digital resources	Preprocessing; lemmatized, and stopwords removal. String similarity, Semantic similarity, generative model	evaluate the document's core concept in terms of how well it embodies basic concepts from	Shallow or weak semantic or textual features were frequently used to compute semantic similarity.

			and shallow feature	related subject domains.	
[36]	Classify Exams question by build a rule by identifying category and assign weight according to Bloom's Taxonomy	85 exams question training set and 62 testing set of questions from Computer Science course from Moratuwa University	Tokenization (Regexptokeni), lemmatization (wordnetlemmatizer), POS (classified BasedTagger), rule based on Path similarity algorithms with lemma similarity	Preprocessing phases comprises the selected tools that provide the appropriate and accuracy format of text for next phases	A high lemma similarity value was used to extract specific question verbs, although this is insufficient for taxonomy categorization. Furthermore, classifying all words as verbs in question increases the possibility of inaccuracy.
[38]	Automatic labeling (weightages)of course questions for certifying their alignment with learning outcomes	A dataset of 150 questions based on the contents of an undergraduate electrical and electronic engineering course was used to train and evaluate machine learning algorithms.	Preprocessing; two phases; labeling questions per 3 levels reduced Bloom's taxonomy transform for machine learning consumption, finally combination of techniques; TF-IDF with ELM performs well	Flexibility of dataset of both training and testing as were retrieved from various source, and combination TF-IDF with extreme learning machine proof to be produced good performance compare with other traditional techniques i.e. SVM	Collapsing 6 levels to 3 levels, may have some limitations in evaluating multi-domain levels as proposed by Bloom's taxonomy. Despite the fact that word weights are usually the same, the nomenclature used to convey them differs. The same sentences could be assigned different weightages.
[34]	Bloom's Taxonomy and ruled based Question analysis approach for measuring the quality of exams papers	Over 900 short essay questions from 30 papers of department of Computing and Information Technology from Sabaragamuwa University were used	Preprocessing; include tokenization, white space removal, and eliminate of non-letter character	Devised six new rule to categorized questions according to the Blooms Taxonomy Levels, the rules go further to categorized the questions into three combined categories. As well as algorithms to check two criteria balanced or unbalanced	It requires a significant amount of manual labor, such as adding rules to each category.
[48]	Document classification using Bag-of-Concepts model from probabilistic knowledge-base	Around 1,503,803 papers were gathered from numerous sources, covering a wide range of themes and fields, such as sports, news, questions and	Preprocessing processes. For categorization and dimension reduction, the Latent semantic analysis (LSA) and LDA based on sklearn used. Other tools include word2vec for pretrained and	Capturing semantic relatedness and conceptual information of words and phrases, as well as higher-level semantics of texts, which is essential for document	The model is only based on concept and word level. To have a deeper knowledge of semantics, a solution on the sentence level is still required.

		answers.	Doc2vec for paragraph learning.	classification.	
[73]	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 most relevant meaning to the phrase and thereby increase contextual entailment.	The techniques utilized for dimensionality reduction are insufficient to reduce the processed data and thereby minimize computational costs.

5. CONCLUSION

The study has reviewed academic articles between 2010 and 2021 which is significant data sample to examine the NLP approaches with their accompanying strengths and challenges in processing educational data such as exam questions, syllabi, and curriculum. The study addresses two analytical approaches utilized in text analysis: statistical and deep learning approaches. Both techniques have strengths and challenges, but deep learning seems to be more effective and accurate than statistical. The common NLP processes for text analysis presented by a number of studies are preprocessing, text representation, classification, and other algorithms depending on the task at hand. The majority of studies modify current techniques or introduce new ones to enhance performance; yet, they face a number of challenges, including: limiting the evidence of the aspect that is more local, i.e., evaluation per regulatory bodies. There are limits to studies conducted on low-resource languages like Swahili, which results in lower performance by misrepresenting the correct meaning of the words in the question. The study has confirmed that the techniques involved in analyzing and evaluating education data have strengths and limitations. Based on a variety of aspects, including the local context, benchmarking educational data per regulatory body criteria, deeper semantic comprehension at the sentence level and above, and computational complexity, further study of NLP techniques is recommended.

REFERENCES

1. Calderon, A., *Massification of higher education revisited.* 2018.
2. Martin, M., *Internal Quality Assurance: Enhancing higher education quality and graduate employability*. 2018: UNESCO.
3. Mok, K.H. 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, 2017. **63**.
4. Leicht, A.H., Julia; Byun, Won Jung, *Issues and trends in Education for Sustainable Development*. UNESCO. 2018, Paris, France: UNESCO Publishing.
5. Laurie, R., et al., *Contributions of education for sustainable development (ESD) to quality education: A synthesis of research.* Journal of Education for Sustainable development, 2016. **10**(2): p. 226-242.
6. Bojorque, R. 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*. 2020. Springer.
7. Gill, S., et al., *Transformative Quality in Higher Education Institutions (HEIs): Conceptualisation, scale development and validation.* Journal of Business Research, 2021. **138**: p. 275-286.
8. Olcay, G.A. and M. Bulu, *Is measuring the knowledge creation of universities possible?: A review of university rankings.* Technological Forecasting and Social Change, 2017. **123**: p. 153-160.
9. Sohel-Uz-Zaman, A.S.M., *Implementing total quality management in education: Compatibility and challenges.* Open Journal of Social Sciences, 2016. **4**(11): p. 207.
10. Kawintiranon, K., et al. *Understanding knowledge areas in curriculum through text mining from course materials.* in *2016 IEEE international conference on teaching, assessment, and learning for engineering (TALE)*. 2016. IEEE.
11. Sanvitha Kasthuriarachchi, K., 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, 2018: p. 179-197.
12. Mohammed, M. and N. Omar, *Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec.* PloS one, 2020. **15**(3): p. e0230442.
13. West, J., *Validating curriculum development using text mining.* The Curriculum Journal, 2016. **28**: p. 1-14.
14. Sangodiah, A., R. Ahmad, and W.F. Wan Ahmad, *A review in feature extraction approach in question classification using Support Vector Machine.* Proceedings - 4th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2014, 2015: p. 536-541.
15. Thalib, I., Widyawan, and I. Soesanti. *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)*. 2020.
16. TAQI, M.K. and R. ALI, *AUTOMATIC QUESTION CLASSIFICATION MODELS FOR COMPUTER PROGRAMMING EXAMINATION: A SYSTEMATIC LITERATURE REVIEW*. Journal of Theoretical & Applied Information Technology, 2016. **93**(2).
17. Chary, M., et al., *A review of natural language processing in medical education.* Western Journal of Emergency Medicine, 2019. **20**(1): p. 78.
18. Silva, V.A., I.I. Bittencourt, and J.C. Maldonado, *Automatic question classifiers: A systematic review*. IEEE Transactions on Learning Technologies, 2018. **12**(4): p. 485-502.
19. Kurdi, G., et al., *A systematic review of automatic question generation for educational purposes.* International Journal of Artificial Intelligence in Education, 2020. **30**: p. 121-204.
20. Ferreira-Mello, R., et al., *Text mining in education.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019. **9**(6): p. e1332.
21. Sangodiah, A., R. Ahmad, and W.F. WAN AHMAD, *TAXONOMY BASED FEATURES IN QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE.* Journal of Theoretical & Applied Information Technology, 2017. **95**(12).
22. Dong, G. and H. Liu, *Feature engineering for machine learning and data analytics*. 2018: CRC Press.
23. Pinto, A., 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)*. 2016. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
24. Inan, E., *SimiT: A Text Similarity Method Using Lexicon and Dependency Representations.* New Generation Computing, 2020. **38**(3): p. 509-530.
25. Sarkar, D. and D. Sarkar, *Feature Engineering for Text Representation.* Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, 2019: p. 201-273.
26. Geigle, C., Q. Mei, and C. Zhai, *Feature engineering for text data*, in *Feature engineering for machine learning and data analytics*. 2018, CRC Press. p. 15-54.
27. Li, H., *Deep learning for natural language processing: advantages and challenges.* National Science Review, 2018. **5**(1): p. 24-26.
28. Allahyari, M., et al., *A brief survey of text mining: Classification, clustering and extraction techniques.* arXiv preprint arXiv:1707.02919, 2017.
29. Huang, W.-J. and C.-L. Liu, *Exploring lexical, syntactic, and semantic features for Chinese textual entailment in NTCIR RITE evaluation tasks.* Soft Computing, 2017. **21**: p. 311-330.
30. Zhang, J. and N.M. El-Gohary, *Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking.* Journal of Computing in Civil Engineering, 2016. **30**(2): p. 04015014.
31. Crain, S.P., et al., *Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond.* Mining text data, 2012: p. 129-161.
32. Abduljabbar, D.A. and N. Omar, *Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination.* Journal of Theoretical and Applied Information Technology, 2015. **78**(3): p. 447.
33. Dhainje, S., et al., *An automatic question paper generation: using bloom's taxonomy.* 2018.
34. Kumara, B., 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, 2019. **5**(1): p. 2-6.
35. Omar, N., et al., *Automated analysis of exam questions according to Bloom's taxonomy*. Procedia-Social and Behavioral Sciences, 2012. **59**: p. 297-303.
36. Jayakodi, K., M. Bandara, and D. Meedeniya. *An automatic classifier for exam questions with WordNet and Cosine similarity*. in *2016 Moratuwa engineering research conference (MERCon)*. 2016. IEEE.
37. Loni, B., et al. *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*. 2011. Springer.
38. Supraja, S., et al., *Toward the Automatic Labeling of Course Questions for Ensuring Their Alignment with Learning Outcomes.* International Educational Data Mining Society, 2017.
39. Aninditya, A., M. Azani 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*. 2019. 112-117.
40. Laddha, M.D., et al., *Classifications of the Summative Assessment for Revised Blooms Taxonomy by using Deep Learning.* arXiv preprint arXiv:2104.08819, 2021.
41. Mohammed, M. and N. Omar, *Question classification based on bloom's taxonomy using enhanced tf-idf.* Int J Adv Sci Eng Inf Technol, 2018. **8**: p. 1679-1685.
42. Pintar, D., et al., *Automatic extraction of learning concepts from exam query repositories.* Journal of Communications Software and Systems, 2018. **14**(4): p. 312-319.
43. Osman, A. 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)*. 2016.
44. Mohamed, O.J., 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, 2019. **14**(2): p. 935-950.
45. Makhlouf, K., et al. *Exam Questions Classification Based on Bloom's Taxonomy: Approaches and Techniques*. in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*. 2020. IEEE.
46. Masua, B. 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, 2020. **33**: p. 106517.
47. Shikali, C.S. and R. Mokhosi, *Enhancing African low-resource languages: Swahili data for language modelling.* Data in brief, 2020. **31**: p. 105951.
48. Li, P., et al., *Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base.* Knowledge-Based Systems, 2020. **193**: p. 105436.
49. Mishra, M., V. Mishra, and H.R. Sharma, *Question Classification using Semantic, Syntactic and Lexical features.* International Journal of Web & Semantic Technology, 2013. **4**.
50. Nguyen, V.-T. and A.-C. Le, *Improving Question Classification by Feature Extraction and Selection.* Indian Journal of Science and Technology, 2016. **9**.
51. Mohasseb, A., M. Bader-El-Den, and E. Haig, *Question categorization and classification using grammar based approach.* Information Processing & Management, 2018. **54**.
52. Mikolov, T., et al., *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems, 2013. **26**.
53. K.M.M Rajashekharaiah, V.B.K., Sreyanka D Somaradder, Dr. P. Suryanarayana Babu *Machine Learning Approach for Automatic Classification of Exam Questions using Blooms Taxonomy and Analysis of Pre-processing Method.* International Journal of Scientific Research in Computer Science Applications and Management Studies 2019. **8**(1).
54. Kim, H.K., H. Kim, and S. Cho, *Bag-of-Concepts: Comprehending Document Representation through Clustering Words in Distributed Representation.* Neurocomputing, 2017. **266**.
55. Bu, F., et al. *Function-based question classification for general QA*. in *Proceedings of the 2010 conference on empirical methods in natural language processing*. 2010.
56. Sarrouti, M., 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)*. 2015. IEEE.

57. Paul, D.V. and J.D. Pawar, *Use of common-word order syntactic similarity metric for evaluating syllabus coverage of a question paper*. 2014.
58. Sherkat, E. 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)*. 2014. IEEE.
59. Cutrone, L. and M. Chang, *Automarking: Automatic Assessment of Open Questions*. 2010. 143-147.
60. Razzaghnoori, M., H. Sajedi, and I.K. Jazani, *Question classification in Persian using word vectors and frequencies.* Cognitive Systems Research, 2018. **47**: p. 16-27.
61. Gomez-Perez, J.M., et al., *Hybrid techniques for knowledge-based NLP*. 2017.
62. Silva, J., et al., *From symbolic to sub-symbolic information in question classification*. Artificial Intelligence Review, 2011. **35**: p. 137-154.
63. Joshi, S., P. Shah, and S. Shah, *Automatic Question Paper Generation, according to Bloom's Taxonomy, by generating questions from text using Natural Language Processing.* 2021.
64. Pota, M., M. Esposito, and G. De Pietro, *A Forward-Selection Algorithm for SVM-Based Question Classification in Cognitive Systems*. 2016. p. 587-598.
65. Zulqarnain, M., et al., *A comparative analysis on question classification task based on deep learning approaches.* PeerJ Comput Sci, 2021. **7**: p. e570.
66. Sultan, M.A., S. Bethard, and T. Sumner. *Towards automatic identification of core concepts in educational resources*. in *IEEE/ACM Joint Conference on Digital Libraries*. 2014. IEEE.
67. Günel, K., et al. *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*. 2014. Springer.
68. Orellana, G., et al. *A text mining methodology to discover syllabi similarities among higher education institutions*. in *2018 International Conference on Information Systems and Computer Science (INCISCOS)*. 2018. IEEE.
69. Badawy, M., et al. *A Text Mining Approach for Automatic Selection of Academic Course Topics based on Course Specifications*. in *2018 14th International Computer Engineering Conference (ICENCO)*. 2018. IEEE.
70. Badawy, M., A. El-Aziz, and H. Hefny, *Exploring and measuring the key performance indicators in higher education institutions.* International Journal of Intelligent Computing and Information Sciences, 2018. **18**(1): p. 37-47.
71. Shaikh, S. and S.M. Doudpotta, *Aspects based opinion mining for teacher and course evaluation.* Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. **3**(1): p. 34-43.
72. Pallegama, P., et al. *Evaluating Teaching Content and Assessments based on Learning Outcomes*. in *2020 2nd International Conference on Advancements in Computing (ICAC)*. 2020. IEEE.
73. Barb, A.S. and N. Kilicay-Ergin, *Applications of Natural Language Techniques to Enhance Curricular Coherence.* Procedia Computer Science, 2020. **168**: p. 88-96.
74. Aalaa Abdulwahab, H.A. and Y.H. Ali, *Documents classification based on deep learning.* Int. J. Sci. Technol. Res, 2020. **9**(02).
75. Zahour, O., et al., *Automatic Classification of Academic and Vocational Guidance Questions using Multiclass Neural Network.* International Journal of Advanced Computer Science and Applications, 2019. **10**.
76. Fromm, H., T. Wambsganss, and M. Söllner, *Towards a taxonomy of text mining features.* 2019.
77. Liu, Y., et al., *Feature Extraction Based on Information Gain and Sequential Pattern for English Question Classification.* IET Software, 2018.

Elia Ahidi Elisante Lukwaro, 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 the Master of Science in Information and Communication Technology at the Open University of Tanzania and the Bachelor Degree with Honor 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 the promotion of 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 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 modeling, agent-based simulation, machine learning, information systems (business informatics), information security, security