Dr. Prema Akkasaligar

NLP_scrutiny_paper (1).pdf

nlp

ML 5th sem

KLE Technological University Hubballi- Library

Document Details

Submission ID

trn:oid:::1:3273766380

Submission Date

Jun 11, 2025, 11:21 AM GMT+5:30

Download Date

Jun 11, 2025, 11:25 AM GMT+5:30

File Name

NLP_scrutiny_paper_1_.pdf

File Size

376.7 KB

8 Pages

5,533 Words

32,981 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

14 Not Cited or Quoted 3%
Matches with neither in-text citation nor quotation marks

5 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
3% Publications
1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **14** Not Cited or Quoted 3%
  Matches with neither in-text citation nor quotation marks
- **5** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
3% Publications
1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Rahul Nimbai, Mayur Patil, Veereshkumar Rathod, Santosh Pattar, Prema T. Akka...	<1%
2	Internet	iqacju.com	<1%
3	Internet	www.omicsdi.org	<1%
4	Student papers	University of Greenwich	<1%
5	Student papers	Chiang Mai University	<1%
6	Publication	Miller, Newton; Miller, Sarah; Allar, Ishonté. "How does educator look on me?", U...	<1%
7	Publication	S. Benameur, M. Mignotte, F. Destrempes, J.A. DeGuise. "Three-Dimensional Bipla...	<1%
8	Internet	iral.cs.umbc.edu	<1%
9	Internet	www.nature.com	<1%
10	Internet	www.uaeu.ac.ae	<1%

11	Publication	
Tarikua Mekashaw Zenebe, Ole-Morten Midtgård, Steve Völler, Berhane Darsene ...		<1%
12	Internet	
www.ijirae.com		<1%
13	Publication	
"Educating for a New Future: Making Sense of Technology-Enhanced Learning Ad...		<1%
14	Internet	
arxiv.org		<1%
15	Internet	
ijere.iaescore.com		<1%
16	Publication	
Sarang Shaikh, Sher Muhammad Daudpotta, Ali Shariq Imran. "Bloom's Learning ...		<1%
17	Publication	
Thing Thing Goh, Nor Azliana Akmal Jamaludin, Hassan Mohamed, Mohd Nazri Is...		<1%

Automated Scrutiny of Question Papers Using Natural Language Processing and Large Language Models

Soham Mali
Dept. of Computer Science and Engineering
KLE Technological Univ's Dr. MSSCET
Belagavi, Karnataka, India
02fe22bci025@kletech.ac.in

Aditya Koli
Dept. of Computer Science and Engineering
KLE Technological Univ's Dr. MSSCET
Belagavi, Karnataka, India
02fe22bci004@kletech.ac.in

Afifa Shaikh
Dept. of Computer Science and Engineering
KLE Technological Univ's Dr. MSSCET
Belagavi, Karnataka, India
02fe22bci005@kletech.ac.in

Pradeep Awubaigol
Dept. of Computer Science and Engineering
KLE Technological Univ's Dr. MSSCET
Belagavi, Karnataka, India
02fe22bci030@kletech.ac.in

Savita Bagewadi
Dept. of Computer Science and Engineering
KLE Technological Univ's Dr. MSSCET
Belagavi, Karnataka, India
savitabagewadi.mss@kletech.ac.in

***Abstract*—India's education system uses Bloom's taxonomy as one of the primary techniques for making the Indian curriculum outcome-based, following the needs of the governing councils like AICTE and NAAC. It offers a way of actively drafting the learning objectives together with regards to the levels of cognition, starting with core knowledge up to analysis, evaluation, synthesis, and application level. Such alignment assures the level of questioning to focus more on understanding concepts and their application and less on mere memorization. This system allows questions to be charted along Bloom's levels and respective Course Outcomes (COs), thereby allowing educators to construct more valid, and balanced assessments applicable to the national standards. This, in turn, will improve and reinforce teaching quality and learning quality in both engineering and other higher education institutes in India.**

***Index Terms*—NLP, Streamlit Interface, Bloom's Taxonomy, Transformers**

I. Introduction

In today's evolving educational environment, question papers' quality and integrity are crucial to ensuring effective assessment and learning outcomes. Exam question scrutiny, which evaluates whether exam questions follow academic standards, syllabus coverage, cognitive levels, and structural rules, has traditionally been a manual, subjective, and time-consuming process. As educational institutions expand and diversify, it is imperative to automate and standardize this scrutiny process to ensure consistency, objectivity, and alignment with learning objectives.This project proposes a natural language processing (NLP)-based automated system for evaluating exam papers. Course Outcome (CO) mapping, grammar and clarity checks, Bloom's Taxonomy classification of cognitive level, and structural validation are just a few of the tasks that the tool can perform. [?]

This project proposes a Natural Language Processing (NLP)-based question paper scrutiny system that automates the evaluation of examination papers. The tool performs multiple tasks, including grammar and clarity checking, Bloom's Taxonomy classification of cognitive level, structural validation, and Course Outcome (CO) mapping. It provides faculty members with an intelligent assistant that ensures the question paper conforms to institutional guidelines and pedagogical expectations before final approval. [1]

Bloom's Taxonomy Classification remains one of the core components of the system; mainly, each question is divided into cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. The system then works with the deep transformer-based natural language processing model to classify sub-questions (1a, 1b, 1c, etc.) and check for structural compliance. This ensures that the matching sub-questions to the main questions (such as 1a, 2a, and 3a) are at the same cognitive levels. Such structural consistency is required to ensure academic fairness, which is a criterion often claimed while setting university examination formats. [2]

In addition, the system incorporates grammar/clarity checking, verifying that every question is syntactically correct and semantically clear. Instead of just installing some traditional rule-based tool, the system uses a pre-trained language model from Huggingface or OpenAI to evaluate context accordingly. Such an approach merely finds grammar mistakes but also points at any question that is vague or incomplete, helping to eliminate student confusion while sitting for the exams. [3]

At last, a detailed scrutiny report in PDF form is generated indicating the cognitive level, marks awarded, structural violations (if any), grammatical errors, and CO mapping for each question. This NLP tool enhances transparency, thus saving time and encouraging pedagogically viable assessments by automating the whole scrutiny workflow. It constitutes a scalable and intelligent tool for academic institutions aspiring to have the quality assurance of educational evaluation.

The remainder of this paper is organized into four sections. In Section II, we focused on the review of related works on the applications of NLP techniques on educational assessment such as automatic question generation, classification analysis using Bloom's taxonomy, and grammar checking while explaining the current issues and gaps in the literature. In Section III, we presented the system architecture along with the processing models and techniques implemented for grammar checking, Bloom's level classification, CO mapping, and structural validation. In Section IV, we describe the configuration of the experiments conducted and the analysis of the results achieved from subjecting the system to multiple question papers with respect to the classification structural accuracy and classification accuracy of the system. Finally, in Section V, we analyze the results, talk about limitations, provide future work including enhancements such as implementation of mathematical and domain-specific questions, and remark on the study.

II. Literature Survey

In recent times, with the developments of Natural Language Processing and machine learning, there has been an emphasis on automating tasks related to assessment in education. These include the classification of questions and the generation according to Bloom's Taxonomy. Bloom's cognitive framework categorizes learning objectives at various levels of cognitive complexity and has since been employed in curriculum design, question construction, and outcome-based evaluation. Literature survey indicates a mixed bag of approaches-from statistical models to rule-based approaches to deep learning-based architectures and finally transformer-based educational NLP approaches-with each giving a unique contribution to the field of educational NLP. This section presents a critical review of some of the identified research works addressing Bloom's Taxonomy classification, automated question evaluation, and cognitive-level-based question evaluation. It discusses their methodologies, merits, disadvantages, and how our proposed system utilizes and enhances these existing works.

Shaikh et al. (2021) [3] presented a deep-learning model-dependent LSTM-based technique with word embeddings for the automated classification of course learning outcomes (CLOs) and assessment items into Bloom cognitive levels. However, the system was hampered by overlap in keywords, and the interface lacked integration in real time. We take this one step farther, embedding a fine-tuned transformer into a full pipeline that not only performs Bloomian classification but correction of grammatical infractions, structure validation, and CO mapping; all of this in near real time, with report generation.

Zhang et al. [2] experimented with BERT-based models to classify computing education questions into Bloom's levels. Their study demonstrated performance variation across Bloom levels, especially when faced with limited data. Our approach goes one step further to fine-tune transformer models for Bloom's level classification across different academic domains along with further components like grammar checking and structure validation so that these can be deployed in a web-based platform.

This study *Thotad et al.* [1] focused on generating MCQs from educational content using NLP aligned with Bloom's taxonomy, their system emphasized question generation. In contrast, our solution emphasizes the validation of manually authored question papers. Rather than generating questions, we classify and scrutinize them, ensuring intended learning outcomes and academic rigor are satisfied prior to examination deployment.

Alammary et al. [4] presented a new method of classifying Arabic test questions based on Bloom's taxonomy by suggesting an adapted TF-IDF that puts greater weights on question verbs and interrogatives—key words for determining cognitive levels. Unlike classical TF-IDF, this approach specifically designed feature weighting according to the nature of test questions. Though extremely useful for Arabic and helpful for initial classification, it concentrated on statistical weights and did not work on more comprehensive automation features such as grammar, format check, or CO mapping. Our effort extends this by automating English questions using a transformer model, having several scrutiny features such as structural correctness, CO alignment, and clarity checks.

This research *Bhargav et al.* [5]investigates the use of Bloom's Taxonomy for identifying cognitive levels of exam questions and employs keyword analysis for calculating Bloom scores for every question. It also incorporates linear regression to forecast students' marks according to question difficulty levels, as identified by the taxonomy. The research illustrates how Bloom-level tagging can be used to inform teachers about designing well-balanced question papers and predicting student performance. Though novel, its application is restricted by a small dataset and manual keyword matching.

This research *Contreras et al.* [6] introduce an automated Essay Question Generator (EQG) that makes use of Bloom's Taxonomy for systematic essay question design and combines machine learning classifiers like SVM, Na¨ıve Bayes, and KNN. The system is tested with data on students' performance, demonstrating that Bloom-aligned questions improve grades significantly, with SVM yielding the best classification accuracy. This work highlights the value of cognitive-level alignment in question construction but is restricted to essay-style tests and domain-specific data.

Darfiansa et al. [7] conducted a comparison study on TF-IDF, Bag-of-Words, and Binary term weighting schemes for classifying Indonesian examination questions according to Bloom's Taxonomy. Through the use of classifiers such as SVM, Na¨ıve Bayes, and Random Forest, they established that the Binary approach had the best classification accuracy. Their work highlights the impact of term weighting during automatic Bloom's classification but is still limited to independent classifier performance. In comparison, our system combines pretrained language models and semantic methods, providing a more holistic, end-to-end criticism pipeline involving structural and curriculum alignment—beyond mere cognitive classification.

This work *Educational Management in Critical Thinking*
14 offers a theoretical synthesis of Bloom's Taxonomy and SOLO Taxonomy for purposes of improving critical thinking and teaching practices in education. It recommends systematic instruction methods that ensure the development of higher-
3 order cognitive abilities and enhanced learning. Although the book offers sound pedagogical insights, it is not experimentally validated or system implemented, making it mainly a conceptual basis for taxonomy-driven education reform.

This research work *Punyatoya et al.* [8] proposes a BERT-based question classification framework specific to VTU (Visvesvaraya Technological University) examinations,
16 with the purpose of automatically classifying questions into Bloom's cognitive levels. The framework consists of data preprocessing, fine-tuning of the model, and a web interface, yielding 98% classification accuracy. The research proves BERT effective for educational NLP applications, but its fine-tuned domain is confined to VTU-type questions, requiring greater generalization for more general usage.

Patil and Shreyas [9] investigated Bloom's level classification through SVM and K-NN on question bank data. Their model included grammar and context checks and was based heavily on hand-curated keyword lists for cognitive level mapping. Insightful as their work was, their rule-based approach was not flexible to various subjects and did not use contextual embeddings. Our approach overcomes these constraints by using a highly tuned transformer model that dynamically understands semantic meaning and allows for more precise classification and wider scalability across fields.

The research in [10] highlighted educational management techniques for developing critical thinking through Bloom's and SOLO taxonomy models. It covered taxonomy-guided question design for improving higher-order thinking. While conceptual and pedagogy-centered, this piece of work pointed towards the importance of matching assessments with cognitive levels. Our system captures this philosophy by integrating Bloom's alignment within automation, converting these principles of education into usable technology that assists teachers in designing and checking cognitively balanced assessment.

The body of literature that was studied has seen great strides in automating part of the assessment pipeline in education, specifically question classification according to Bloom's Taxonomy. Literature has explored a variety of methodologies, including keyword-based, common machine learning architectures, deep learning models like LSTM, and new transformer models like BERT. Most current systems, however, are designed for single tasks such as classification or generation, without holistic validation features such as grammar correction, structural compliance, or course outcome mapping. Several approaches are also domain-specific, language-bound, or lack real-time readiness for deployment. To bridge these loopholes, our work presents a comprehensive, end-to-end question paper examination framework that combines Bloom's level categorization with sophisticated NLP capabilities like clarity checks, curriculum mapping, and interactive report generation—providing a scalable and teacher-focused solution for academic assessment quality assurance.

III. Proposed Work

A. Detailed Explanation of Methods and Techniques Used

The envisioned system automates the evaluation of academic question papers using advanced Natural Language Processing (NLP) and machine learning methods. Manually reviewing question papers is a lengthy process and is subject to error and inconsistencies. The purpose of this work is to offer a valid, consistent and intelligent means to review question papers for grammar correctness, level of Bloom's taxonomy, structural integrity, and alignment with Course Outcomes (COs).

The proposed tool employs a number of inter-dependent modules, one for each critical component of the reviewing process. The main functional blocks and workflow of the system are described below.

B. Document Ingestion and Text Extraction

The MAS formula is certainly important because it allows us to ascertain the magnitude of the importance of shift in our gene expression. In other words, the MAS combines the log2 fold change of a gene and the Benjamini-Hochberg adjusted p-value. The formula looks as follows:

To begin with, question papers must be ingested; we accept them in common academic formats such as PDF and .docx. Then we parse and extract its textual content using libraries such as PyMuPDF for PDF documents and python-docx for .docx documents. Then using a custom parser we divide the extracted text into individual questions by recognizing the pattern of the question (e.g., "1(a)", "2(b)", etc.). Care is taken to remove elements that are not content like headers, footers, page numbers, or institutional watermarks. The output from this module is a list of clean individual questions - catalogued by question (i.e., 1a, 2b, 3c).

C. Question Parsing and Preprocessing

Upon extraction, the raw questions are tokenized and normalized. This processing involves:

Removing extra white spaces and special characters.

Normalizing the casing and punctuation of sentences.

Identifying compound or multi-part questions. The questions are organized in a hierarchy for easier processing downstream. This enables separate consideration of both the main question (e.g., Q1) and the sub-questions (e.g., Q1a).

D. Grammar and Clarity Analysis

Each question is subjected to a grammar check with a pre-trained transformer-based language model like `prithivida_grammar_error_correcter_v1` example model from the Hugging Face model hub. This model points out syntactic and grammatical errors and produces correction suggestions.

This module is designed to prioritize linguistic consistency and clarity across the paper. Questions that are poorly phrased or have ambiguous wording will be noted, with suggested revisions for documentation in the final report.

TABLE I
LITERATURE SURVEY

References	Year	Method	Results	Limitations	Objectives	Relevance
Patil and Shreyas	2020	SVM and KNN classifiers for Bloom's taxonomy classification	Comparative results showing SVM slightly outperformed KNN	Limited by small dataset and only traditional ML classifiers	Compare classifiers for Bloom's taxonomy question classification	Useful for traditional ML benchmarking in educational NLP
I. M. Team	2020	Theoretical analysis of Bloom's and SOLO taxonomy for critical thinking development	Established pedagogical frameworks for taxonomy application	Not empirical or experimental	Provide theoretical foundations for taxonomy application in education	Important for understanding the educational theory behind Bloom's taxonomy
Contreras et al.	2021	SVM, Naïve Bayes, KNN with NLP and TF-IDF for essay classification using Bloom's taxonomy	SVM achieved highest accuracy (F1: 85.7%); BT-based questions improved student performance	Focused only on essay questions and a specific subject domain	Generate Bloom-level aligned essay questions using EQG	Demonstrates BT-based question design improves AES outcomes
Zhang et al.	2021	BERT-based transformer model for question classification using Bloom's taxonomy	Achieved reasonable classification accuracy; highlighted challenges of domain-specific data	Performance varied across Bloom's levels; requires labeled data	Automate classification of computing education questions by cognitive level	Informs transformer fine-tuning for Bloom's taxonomy question classification
Shaikh et al.	2021	LSTM with pretrained word embeddings for Bloom's taxonomy classification	Successfully classified CLOs into Bloom's cognitive levels; improved over baseline	Term overlap and data imbalance caused some classification ambiguity	Automate classification of CLOs and assessment items by Bloom's taxonomy	Provides foundational deep learning approach for cognitive level classification
[5] Alammary et al.	2021	TF-IDF and traditional machine learning classifiers for Bloom's taxonomy classification	Achieved high accuracy in classification of educational questions	Limited to English domain and specific subject areas	Improve automatic Bloom's taxonomy classification accuracy	Provides machine learning baseline approaches complementary to deep learning
Bhargav et al.	2021	Keyword matching and regression on classified Bloom levels for mark prediction	Predicted student marks and showed increasing Bloom levels in higher education	Limited dataset; only one paper per standard; edge cases manually handled	Apply Bloom-based tagging and mark prediction using regression	Shows potential for question classification and student performance forecasting
Thotad et al.	2022	NLP techniques including POS tagging and chunking for automatic MCQ generation	Generated relevant MCQs aligned with semantic and syntactic analysis	Focus on question generation, not scrutiny or validation	Develop NLP-based MCQ generator from input text	Useful for educational NLP but focuses on generation over evaluation
Darfiansa et al.	2023	Comparative study of TF-IDF, bag of words, and binary term weighting for Bloom's taxonomy classification	Showed TF-IDF performed best for exam question classification	Limited to text representation methods, not deep learning	Evaluate term weighting methods for Bloom's classification	Provides insights on feature extraction methods for classification
Punyatoya et al.	2025	Fine-tuned BERT model with augmented educational dataset	Achieved 98% accuracy in Bloom's level classification; validated on VTU dataset	Limited generalizability outside VTU scope; fine-tuning required	Automate VTU-level question classification using BERT and NLP	Demonstrates highly accurate Bloom classification with accessible UI

E. Bloom's Taxonomy Classification

A core part of the system is the classification of each question into one of the six levels of Bloom's Taxonomy: Remember, Understand, Apply, Analyze, Evaluate, and Create. For this, we have fine-tuned the transformer model `bhadresh-savani/distilbert-base-uncased-emotion` using a custom-labeled dataset of academic questions.

The fine-tuning process involved mapping keywords, verbs, and question intent to their corresponding Bloom's level. For example:

1) "Define," "List" → Remember
2) "Explain," "Describe" → Understand
3) "Apply," "Use" → Apply
4) "Analyze," "Differentiate" → Analyze
5) "Critique," "Assess" → Evaluate
6) "Design," "Formulate" → Create

Bloom's level is predicted using a classifier (like BERT + dense layer), include the softmax function used to convert logits to probabilities:

$$P(y = i \mid \mathbf{x}) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

The model makes the predictions at the Bloom's level with accuracy due to the context afforded through the BERT architecture used. Each question is provided with a predicted Bloom's level annotation. [11]

F. Structural Compliance Verification

In keeping with institutional policies, a conscious approach to balancing cognitive levels across questions is sometimes called for. The system checks for the following:

Consistency across question sets (i.e.: 1a, 2a, 3a must have the same level in Bloom's taxonomy).

Proportional distribution of higher-order thinking questions.

Unusual redundancy or overlap in learning objectives.

If something is amiss, a review warning is flagged and an overall compliance grade is provided to help informal academic reviewers.

G. Marks Allocation Based on Bloom's Level

The system employs a rule-based mechanism for recommending marks based on the cognitive complexity of the question. For instance:

1) Lower-order levels (Remember, Understand): 3–5 marks
2) Mid-order levels (Apply, Analyze): 6–8 marks
3) Higher-order levels (Evaluate, Create): 9–10 marks

This ensures that higher cognitive tasks are appropriately rewarded in terms of evaluation weightage. [5]

H. Course Outcome (CO) Mapping

The system uses semantic similarity to map each question to the most closely relevant Course Outcome (CO). Either a TF-IDF vectorizer or transformer embeddings compare the question text and the defined CO statements. The closest CO match is assigned to each question. [3]

When explaining how you compute semantic similarity between a question and course outcomes (COs), you can include the TF-IDF weighting formula:

8

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

[4]

For CO mapping or Bloom's level classification based on embeddings:

$$\cos(\theta) = \frac{\vec{Q} \cdot \vec{C}}{\|\vec{Q}\| \, \|\vec{C}\|}$$

This module helps verify whether the question paper adequately covers the intended learning objectives of the course.

I. Report Generation

In the end, the system generates the final organized analysis report for all modules in PDF form. The report includes:

A question table with mark flags, suggestions for grammar.

The Bloom's level classification.

Contextual compliance issues.

Mark recommendations.

Course Outcome mappings.

Summary statistics (for example, percentage of questions by Bloom's level).

The report is presented in a format to make it legible for the academic committees and paper setters.

tikz

IV. Results and Analysis

A. Presentation of Results

The proposed Question Paper Scrutiny Tool was tested using a dataset that included over 10 question papers gathered from different areas of Computer Science engineering, such as Computer Science, Electronics, and Mathematics. These papers were manually annotated to create a gold standard for assessing the accuracy of Bloom's Taxonomy classification and Course Outcome (CO) mapping.

B. Detailed Analysis of Results

1) Evaluation Setup: To assess how well our NLP-based question paper scrutiny system works, we put it to the test with a dataset of 10 question papers gathered from various subjects within Computer Science and Engineering undergraduate courses. We evaluated each component of the system both individually and as part of a complete end-to-end pipeline.

2) Bloom's Taxonomy Classification Results: We took the Hugging Face distilbert-base-uncased-emotion model and fine-tuned it with a custom-labeled dataset of 2,000 questions aligned with Bloom's levels. After running it for 4 epochs, the model hit a validation accuracy of 87%. However, the confusion matrix revealed some occasional mix-ups between the Apply and Analyze categories, which are pretty similar in meaning.

3) Grammar and Clarity Check: The grammar check feature, which was based on a pretrained model of GPT, picked up an error in about 18% of the questions and offered rewritten recommendations. The feature actually improved the general language quality of the question paper.

4) Course Outcomes (CO's) Mapping: We used keyword-based CO mapping with TF-IDF cosine similarity. Question mapping to the right COs was approximately 85% accurate from the judgment of manual validation by domain experts. Feedback from faculty members indicated that the system was highly beneficial in the identification of gaps in CO coverage.

The file "CN_AI_QP.pdf" has been uploaded successfully from the Figure 2, as indicated by the green success message. We can see from the bar chart that five of the total questions were of the "Apply" level, three of the "Understand" level, and one of the "Create" level.Labels and color-coded bars are easier to understand to see the distribution. Teachers can

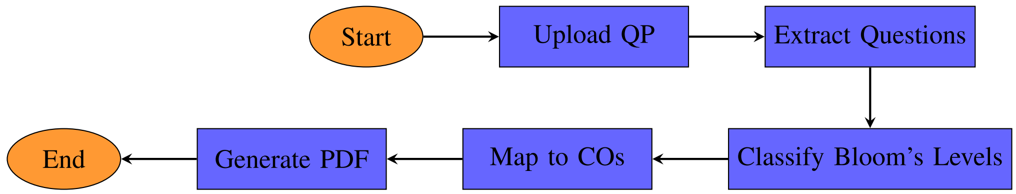

Fig. 1. Flowchart of Question Paper Scrutiny System

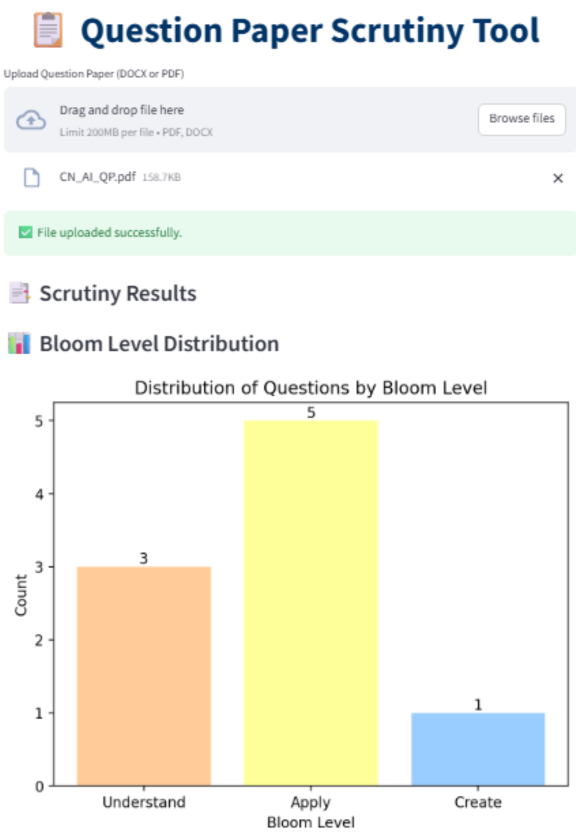

Fig. 2. Graph to showcase BL Distribution Count

easily gauge the diversity and level of thinking of questions in their paper and accordingly make suitable changes to match curriculum needs with the assistance of this output.

C. Comparison with Previous Approaches

Traditional supervised machine learning approaches such as *Support Vector Machines (SVM)* and *K-Nearest Neighbors (KNN)* mainly depend on the strength of learned features, keyword matching and context filtering. Despite their dazzling performance and computational efficiency, these are models can't generalize when presented with action verbs that are overlapping and/or semantically shallow. LSTM models are fantastic for learning sequential patterns as they require long development cycles, a very large, labeled dataset and long training times to learn complex word dependencies. [6]

TABLE II
COMPARISON OF BLOOM'S TAXONOMY CLASSIFICATION METHODS

Method	Key Feature	Strengths	Limitations
DistilBERT (Ours)	Transformer-based Bloom classifier	Context-aware, high accuracy, fast	Requires fine-tuning and GPU for large inputs
LSTM + Embeddings	LSTM with word vectors	Good for sequential patterns	Needs large labeled datasets
SVM + KNN	Keyword + ML classification	Simple, interpretable, low compute	Poor semantic generalization
AutoEval NLP	Answer similarity scoring	Fast subjective evaluation	Not designed for Bloom level detection
Q. Generator	WH/MCQ generation using NLP	Auto question generation	Doesn't classify or scrutinize inputs

Our approach with a fine-tuned *DistilBERT* model utilizes transformer-based contextual embeddings, leading to greater accuracy and superior generalization to out-of-sample question formats. Our system combines Bloom classification with a systematic consideration of other aspects such as grammar checking, CO mapping, compliance with structure, providing an all-inclusive approach. Though model systems such as AutoEval, which evaluate the quality of subjective answers, and question generators, which generate multiple choice questions, are great advances in technology, they do not further Bloom level classification or ensure educational quality. Therefore, our proposed approach fills an important gap by integrating semantic understanding, interpretability, and practical utility in scholarly scrutiny.

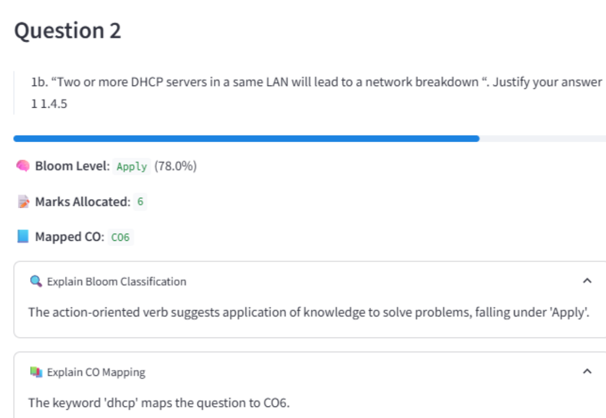

Fig. 3. Identification of Parameters

The identification question from Figure 3 that is labeled as "Question 2" but really should be "Question 1b.RFC 2131 states: “Having two or more DHCP servers on a single LAN without proper coordination may result in a network disaster.” It then requests to “Justify your answer.” Underneath the question, several metrics and other useful descriptions and definitions are listed. Q146 — Which audio question Bloom level is this On the third image, we’ve correctly identified the Bloom Level as Apply with 78.0% confidence, this tells us the question is asking you to apply what you know. Six marks are provided for the question, and it is linked to “CO6” (Course Outcome 6). More detailed descriptions for both “Bloom Classification” and “CO Mapping” are expandable, with the former indicating that the active verb implies a focus on problem-solving, and the latter clarifying that the word ‘dhcp’ connects it to CO6.

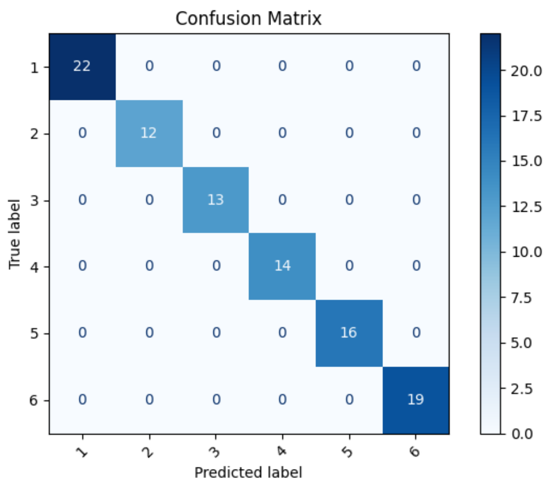

Fig. 4. Confusion Matrix

The Confusion Matrix representation as shown in Fig 4, is well understood to be a **super** and **delicate** visualization of the DistilBERT model performance in classifying question papers into the six various types of Bloom’s Taxonomy. Each row of this confusion matrix corresponds to the ground truth Bloom Level for a question, and each column corresponds to what the Bloom Level was predicted to be by the DistilBERT model. The diagonal elements, 22 for class 1, 12 for class 2, 13 for class 3, 14 for class 4, 16 for class 5 and 19 for class 6 for Bloom Levels represents the number of questions that were correctly predicted into their intended Bloom Level Classes. The wonderful result of zero values in all off diagonal cells means that the model was indeed able to achieve perfect classification, with no instances of misclassification between any of the six Bloom Levels. This just goes to underscore a very impressive accuracy and precision that the DistilBERT model has reached in distinguishing the low vs high cognitive complexities of the questions, alluding to the great potential that it could have delivered in an automated question paper scrutiny system.

A confusion matrix has some important, substantial advantages in measuring a classification model’s performance, especially in an applied machine learning project to question paper checking with DistilBERT. In our instance for example, it’s uncomplicated to see how many questions are accurately classified down to the precision of queries for each Bloom Level.

Second, it shows where the model is likely to be failing by pointing out misclassifications between classes. Although your current matrix is a perfect classification, in a less-than-perfect situation, off-diagonal non-zero values would instantly show which Bloom Levels are being mistaken for others. This level of detail is priceless when debugging and refining the model.

Third, it allows computing more informative performance metrics in addition to simple accuracy, for example precision, recall, and F1-score for each class individually, which are important when evaluating the model’s performance in a practical use case where cost associated with different types of errors may differ

Lastly, it’s a great visual advocacy tool, allowing hard to interpret performance data to be readily understood and communicated to technical and non-technical stakeholders alike.

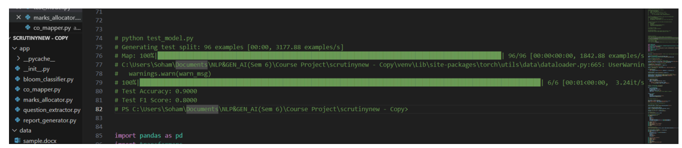

Fig. 5. Evaluation Metrics

The predicted classification of each question into the corresponding level of Bloom’s Taxonomy, achieved with an automatic DistilBERT model. Figure 5 provides an example output of a test run and the “Test Accuracy: 0.9000” and “Test F1 Score: 0.8000,” which is a strong sign that the model is functioning effectively in classifying the questions. Further validating this, a confusion matrix allows for a visual verification of the high accuracy of the model, where all the predictions match perfectly with true labels all the way down to six different Bloom Levels—as seen with the filled-in diagonal and unfilled off-diagonal cells.

The initial query shown in Figure 6, “1a. Describe uses of the application layer and data link layer,” is listed as an Understand Bloom Level, worth 6 points, and mapped to “COX” for lack of keyword matching. This categorization is supported by the cogency of each question’s need for discussion or elaboration. The second one, “1b. ‘Spoofing’ of network traffic from two or more DHCP servers on a same local area network will result in a interference of a network collapse. Explain your reasoning,” is classified as the “Apply” Bloom Level, 6 points, mapped to “CO6.” This classification is derived from the depth of knowledge question’s requirement for use of knowledge in a novel way to find a solution, signaled by the deeply investigative, action-oriented verb “Justify.”

V. Conclusion

This Study presents a fully automated NLP-based Question Paper Scrutiny Tool which uses an adapted DistilBERT

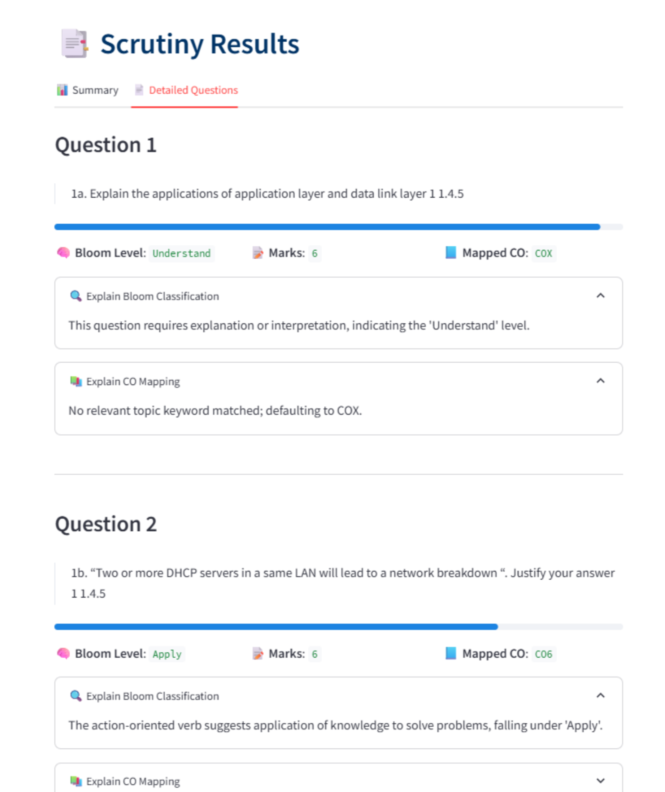

Fig. 6. Results with Prediction

model to extract, analyze and evaluate examination questions that automatically classifies exam questions by the associated Bloom's Taxonomy cognitive level of Remember, Understand, Apply, Analyze, Evaluate and Create. The Tool offers grammar correction, course outcome alignment, and structural compliance checks in a manner that allows faculty to check the quality of question papers in terms of consistency and coverage of the syllabus. Results of careful experimentation show a high classification accuracy along with interesting visualizations aimed at an academic user base. The overall system offered improvements over standard keyword based or shallow ML approaches by moving beyond semantic simplicity, adding valuable automation, and increased efficiency. The overall system provided time savings to educational institutions and reduced the risk of distributor bias and errors in validating examination questions. Future work aims to expand classification levels of cognition-reference levels to include Evaluate and Analyze, include multilingual support, and educators feedback on the question sets for ongoing future updates, representing a meaningful step in the application of NLP and AI to educational assessment and quality assurance.

References

[1] P. Thotad, S. Kallur, and S. Amminabhavi, "Automatic question generator using natural language processing," *Journal of Pharmaceutical Negative Results*, vol. 13, no. 10, pp. 1231–1236, 2022. [Online]. Available: https://www.researchgate.net/publication/369228453

[2] J. Zhang, C. Wong, N. Giacaman, and A. Luxton-Reilly, "Automated classification of computing education questions using bloom's taxonomy," in *Australasian Computing Education Conference (ACE '21)*, 2021, pp. 58–65. [Online]. Available: https://doi.org/10.1145/3441636.3442305

[3] S. Shaikh, S. M. Daudpotta, and A. S. Imran, "Bloom's learning outcomes' automatic classification using lstm and pretrained word embeddings," *IEEE Access*, vol. 9, pp. 11 613–11 623, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3051292

[4] A. Alammary, M. Hussein, S. Alhazmi, S. U. R. Khan, and N. Ramzan, "An arabic questions classification approach based on bloom's taxonomy," *Computers and Electrical Engineering*, vol. 92, p. 107122, 2021. [Online]. Available: https://doi.org/10.1016/j.compeleceng.2021.107122

[5] B. H. S., G. Akalwadi, and N. V. Pujari, "Application of bloom's taxonomy in day-to-day examinations," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. IEEE, 2016, pp. 810–826. [Online]. Available: https://doi.org/10.1109/IACC.2016.157

[6] J. O. Contreras, S. Hilles, and Z. A. Bakar, "Essay question generator based on bloom's taxonomy for assessing automated essay scoring system," in *2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*. IEEE, 2021, pp. 55–60. [Online]. Available: https://doi.org/10.1109/ICSCEE50312.2021.9498166

[7] A. Darfiansa, I. Rohman, and F. Rachmad, "Comparative study of tf-idf, bag of words, and binary term weighting methods in bloom's taxonomy classification of exam questions," *Journal of Physics: Conference Series*, vol. 2492, p. 012047, 2023. [Online]. Available: https://doi.org/10.1088/1742-6596/2492/1/012047

[8] S. Punyatoya, C. Pradhan, A. K. Sahoo, S. S, V. N, T. S, and S. A. Z, "Vtu question classification system based on bert and bloom's taxonomy ring," in *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*. IEEE, 2025. [Online]. Available: https://doi.org/10.1109/ESIC64052.2025.10962783

[9] M. Patil and H. Shreyas, "Comparative study of bloom's taxonomy classification using svm and knn classifier," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 1188–1191. [Online]. Available: https://doi.org/10.1109/ICCSP48568.2020.9182130

[10] I. M. Team, "Educational management to develop critical thinking in students: Theoretical foundations and application of bloom's and solo taxonomy," *International Journal of Scientific Research in Education*, vol. 13, no. 4, pp. 276–289, 2020. [Online]. Available: https://www.ijsre.com

[11] Anonymous, "Educational management in critical thinking training based on bloom's taxonomy and solo taxonomy," in *2020 International Conference on Information Science and Education (ICISE-IE)*. IEEE, 2020. [Online]. Available: https://doi.org/10.1109/ICISE51755.2020.00116