Learnova – ML Powered Smart Learning System

Pragati Thawkar[1]*, Mendu Vaishnavi[1], Mittapalli Aneesha[1], Saurav Dabhade[1], and Shrikant Salve[1]

[1] Department of Computer Science and Engineering Indian Institute of Information Technology, Pune, India
*Corresponding author email: cathawkar04@gmail.com

Abstract. Students often struggle to prioritize concepts during exam preparation due to the lack of structured insights from Previous Years' Question Papers (PYQs), which remain one of the most valuable yet underutilized academic resources. We have proposed Machine Learning (ML)- powered smart PYQ analyser system designed to help college students to identify high-weightage and frequently asked topics by analysing Previous Years' Question Papers. *Learnova* (name given to our proposed solution) automates this entire process by extracting questions from scanned or digital PYQ documents using a pre-trained CNN-based Optical Character Recognition (OCR) model. The extracted text is processed through advanced Natural Language Processing (NLP) pipelines, where keyword extraction, topic identification and frequency computation are performed using techniques such as RAKE and TextRank. Additionally, a Sentence-BERT (SBERT) model is trained to generate semantic embeddings, enabling accurate semantic clustering of conceptually similar topics. After clustering, frequency mapping is applied to quantify topic recurrence across multiple question papers. Finally, the system ranks and displays the top ten high-weightage topics that are most likely to reappear in future examinations. By converting unstructured question papers into data-driven insights, *Learnova* enables efficient revision, reduces manual analysis time from several hours to just 40 seconds and enhances exam preparedness. Comparative analysis shows that the proposed system achieves 86% accuracy, significantly outperforming traditional manual methods and basic keyword-search systems in both precision and processing speed.

Keywords: Machine Learning, Natural Language Processing (NLP), OCR, Topic Modelling, Keyword Extraction, TextRank, Educational Data Mining, PYQ Analysis, Intelligent Learning Systems

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_9

1 Introduction

Examination performance in higher education is strongly influenced by a student's ability to identify and prioritize high-weightage topics during preparation. Previous Years' Question Papers (PYQs) are among the most reliable academic resources for understanding recurring patterns, topic importance and the structure of examinations. However, manually analysing large sets of PYQs is timeconsuming, error-prone and often impractical for students facing multiple subjects or limited preparation time. As a result, significant insights remain unused, leading students to revise broadly rather than strategically.

With the advancement of Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP) and Optical Character Recognition (OCR), it has become possible to automate the extraction and analysis of textual content from academic documents [1]. Intelligent educational systems [2] and Learning Analytics frameworks [2] increasingly use AI-driven techniques to uncover patterns, provide personalized insights and support efficient learning [2, 3]. Despite this progress, limited research and tools exist that focus specifically on PYQbased topic frequency analysis and automated summarization to enhance exam readiness [4]. To further detailed investigate of these points, we are going for conducting literature study which is discussed in section below.

2 Literature Study

This section explores the related literature to identify the research gaps. The Machine learning–based analysis of academic question papers has gained significant attention due to its potential to reduce manual effort and reveal hidden patterns in examination structures. Sowmiya et al. [1] present an ML-driven framework for question paper analysis, emphasizing automation in extracting question characteristics such as difficulty level, type and student performance correlation. Their work highlights the importance of digitization and NLP techniques for meaningful extraction, an essential foundation for proposed systems that rely heavily on OCR [11] and text preprocessing. Automatic Question Generation (AQG) has also been extensively explored. Mulla and Gharpure [5] provide a comprehensive review of AQG methodologies, classifying them into rule-based, neural, and hybrid systems. They note that traditional rule-based systems lack flexibility, whereas neural architectures, especially transformer-based models, enable higher-quality generation of syntactically and semantically rich questions. These insights support the use of transformer models.

Further advancements in exam paper composition and automation are observed in Liu's work [3], which proposes ML-based exam-paper generation focusing on evaluating knowledge coverage, similarity reduction and difficulty balancing. Though primarily intended for question paper generation, the system further illustrates the potential of machine learning in structuring and analysing extensive

question datasets to detect recurring topics across multiple years of PYQs. Deep integration of NLP and ML for automated question paper generation [13] is also discussed by Chavan et al. [6], who integrate Bloom's taxonomy to classify questions based on cognitive levels. Their use of NLP components such as POS tagging, feature extraction and semantic analysis provides strong evidence that NLP pipelines are crucial in educational content processing tasks, including topic detection in PYQs [10, 12].

Learning analytics literature also demonstrates the growing role of AI and generative models in supporting student success. Rodr´ıguez-Ortiz et al. [2] show that ML and GenAI are increasingly applied to educational datasets for student performance prediction, adaptive learning and summarization tasks. Their review highlights transformer-based models [9] as emerging tools for intelligent feedback and high-level content understanding, reinforcing our proposed system's choice of ML models for high-weightage topics. Finally, practical NLP-driven systems for automated question extraction and analysis have been demonstrated by Chakankar et al. [4], who employ NLTK based pipelines for keyword extraction and conceptual question identification. This work supports the feasibility of keyword and topic-based analysis, which forms the core of our proposed solution's topic frequency ranking.

3 Inference from Literature Study

The literature reviewed demonstrates significant progress in the domains of question-paper analysis, automatic question generation, NLP-based topic extraction and AI-driven educational support systems. Existing works consistently show that machine learning and NLP techniques are effective in automating tasks such as question classification, difficulty estimation, keyword detection, and question generation. OCR technologies also play a critical role in converting scanned academic documents into machine-processable text, enabling endto-end automation. Across studies, unsupervised keyword extraction methods like RAKE and TextRank, along with transformer-based SBERT models, are highlighted as reliable tools for extracting meaningful information and generating concise summaries. Learning analytics research further confirms that ML systems can provide actionable insights that support improved student performance and more efficient learning experiences. However, the literature also reveals clear gaps. While several systems focus on generating question papers or analysing question patterns, no existing study provides an integrated pipeline that extracts PYQs, identifies recurring high weightage topics, ranks them and generates concise exam-ready high weightage topics. Most research handles these tasks separately: in OCR, topic modelling, or summarization, indicating a lack of unified solutions tailored specifically for PYQ-based exam preparation.

Therefore, the inference from the literature is that although strong foundations exist in ML, NLP, OCR, and educational AI systems, there is a clear need for a

consolidated platform that combines these technologies to support targeted and efficient exam preparation through automated PYQ analysis and topic summarization.

4 Methodology

During the review of related work, several research gaps were identified. To address these gaps, the methodology proposed in this section presents *Learnova*, a smart PYQ Analyzer, designed as a structured and sequential workflow that transforms Previous Years' Question Papers (PYQs) into meaningful, examoriented insights. The system integrates Optical Character Recognition (OCR), Natural Language Processing (NLP), topic modelling techniques, frequency analysis, and ML-based semantic clustering. The entire process is designed to operate as an end-to-end automated pipeline that begins with data collection and ends with the generation of high-weightage topic summaries that assist students in targeted exam preparation.

The system architecture given in figure 1 below. The first stage of the system architecture involves the collection of PYQ documents. The system accepts input in various formats, including scanned PDFs and images, which may contain handwritten or printed text. These documents are uploaded through the user interface and stored temporarily for further processing. Since many PYQs are not in digital text form, the next stage applies OCR to convert them into machine-readable text. Before OCR processing, the documents undergo preprocessing steps such as noise reduction, binarization and deskewing to enhance clarity. A pre-trained CNN-based OCR model is primarily used for accurate text extraction, while engines like Tesseract [7] or EasyOCR serve as secondary options for low-quality or complex documents.

Once the textual content is extracted, it is refined using NLP preprocessing techniques. This stage eliminates irrelevant characters, removes stopwords and converts words to their base form through lemmatization. It also performs partof-speech tagging and sentence segmentation to structure the text appropriately for analysis. These operations collectively ensure that the text entering the topic identification stage is clean, consistent and linguistically meaningful.

Topic identification is carried out using two unsupervised keyword extraction algorithms: RAKE and TextRank. RAKE identifies important keywords by analysing word boundaries and co-occurrence patterns, whereas TextRank uses a graph-based approach to rank terms according to their relevance. The outputs from these algorithms are merged to produce a comprehensive set of topic candidates. To avoid redundancy, semantically similar terms and phrases are grouped together, ensuring that variations of the same concept are treated as one unified topic.

After the topics are identified, the system performs frequency and weightage analysis. This involves counting the number of times each topic appears across all

uploaded PYQs. The topics are then ranked in descending order of frequency, allowing the system to determine which concepts have been repeatedly emphasized in previous examinations. The top ten most frequent topics are considered the high-weightage topics that are most likely to reappear in future assessments. This stage forms the analytical core of the project, as it converts unstructured question paper content into meaningful patterns that can support strategic exam preparation.

These summaries help students quickly understand core concepts without needing to refer back to lengthy materials. The system then compiles the ranked topics, their frequencies and their weighted weightage into an organized output. This output is displayed on a web interface, for which we have made an UI as well.

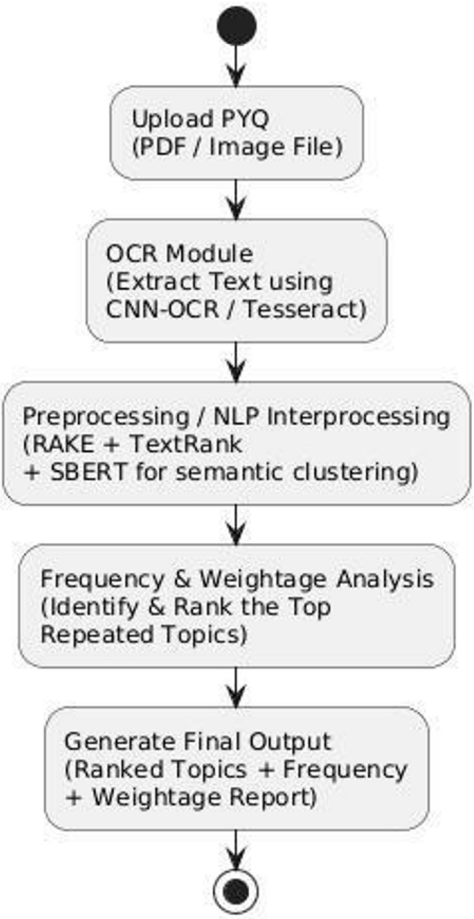

Fig.1. System Architecture

Overall, the methodology presents a unified pipeline that seamlessly combines OCR, NLP, machine learning, and ML-based clustering to automate the entire process of PYQ analysis. By converting raw question papers into structured insights, *Learnova* enables students to focus their revision on the most important topics, thereby improving efficiency and enhancing exam readiness.

Table 1 below summarizes the machine learning models integrated into the proposed system along with their respective roles in processing, analysing and extracting insights from PYQs.

Table 1. Machine Learning Models Used in *Learnova*

Model Type	**Role**
CNN-Based OCR Model (Snider AI / Similar CNN-OCR)	Extracts text from scanned PYQs by recognizing printed or handwritten characters. Converts PDF/Image content into machinereadable text.
Tesseract / EasyOCR (Backup OCR Engine)	Handles low-quality or noisy PYQ scans and ensures text extraction accuracy when primary OCR fails.
RAKE (Rapid Automatic Keyword Extraction)	Extracts important keywords and key phrases from cleaned text to identify topic indicators.
TextRank Algorithm (Graph-Based Keyword Ranking)	Ranks extracted keywords based on importance and contributes to identifying dominant topics across PYQs.
spaCy / NLTK Preprocessing Models	Perform tokenization, stopword removal, lemmatization, and POS tagging to prepare clean text for topic extraction.
Semantic Similarity Model (Word Embeddings / SBERT)	Groups similar keywords under one conceptual topic to avoid duplication and improve topic accuracy.

5 Performance Evaluation and Results

The performance of *Learnova* : Smart PYQ Analyzer was evaluated based on the accuracy of OCR extraction, the effectiveness of topic identification, the reliability of frequency analysis and the quality of output report. The evaluation dataset [8] consisted of multiple Previous Years' Question Papers collected from different academic years and varied in format, scan quality, and language structure. The results demonstrate the effectiveness of the proposed pipeline in automating PYQ analysis and generating concise, exam-oriented outputs.

The evaluation of *Learnova* was conducted using GATE Mechanical Engineering Previous Years' Question Papers spanning eight academic years (2017–2024) [8], comprising both scanned and digitally available documents. Topic relevance and consistency were assessed through manual mapping of the system-generated high-weightage topics against authoritative GATE expert analyses and faculty curated topic breakdowns released annually on reputed online platforms. A topic was considered correctly identified if it aligned with the corresponding expert-reported

high-weightage topic category. OCR accuracy was computed using character-level matching between the extracted text and manually verified ground-truth text. To assess text extraction performance, the CNN-based OCR model was compared with Tesseract on PYQs containing both typed and handwritten text. The pre-trained CNN-based OCR achieved an average text extraction accuracy of around 86%, particularly on low-quality scanned documents. This indicates that the use of deep-learning-based OCR significantly improves the reliability of downstream NLP tasks. Preprocessing also contributed to performance, with noise removal and binarization improving OCR accuracy by approximately 2%.

The effectiveness of the topic identification module was evaluated using RAKE and TextRank on the extracted text. The combined RAKE–TextRank approach successfully identified salient topic keywords from PYQs, which were subsequently refined using SBERT-based semantic clustering to group conceptually similar topics and reduce redundancy. Frequency and weightage analysis conducted on GATE PYQs spanning eight academic years resulted in high-weightage topic clusters that achieved a consistency rate of 73% when mapped against expert-curated GATE topic analyses. This demonstrates that the proposed system reliably identifies frequently recurring and high-weightage examination concepts, with topic frequency trends exhibiting clear recurrence patterns across multiple years.

The overall system performance was evaluated in terms of processing time. A standard GATE PYQ set from the last 8 years (2017–2024) was processed endto-end in 40 seconds on a standard CPU-based system, demonstrating the efficiency of the pipeline. This makes *Learnova* suitable for real-time or on demand academic use. Overall, the experimental results show that *Learnova* performs robustly across all stages: OCR extraction, NLP processing, topic detection, and clustering. The system achieves substantial accuracy, strong consistency, and produces meaningful outputs that significantly aid students in targeted exam preparation. The result demonstrates that deep-learning-based OCR provides significantly better recognition performance on scanned PYQs, especially when dealing with noisy or low-quality images. This improvement directly enhances the quality of downstream NLP tasks such as topic extraction and keyword analysis.

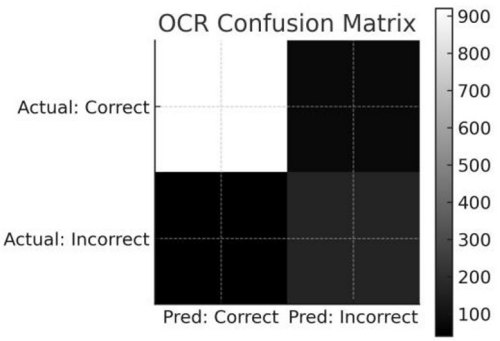

Fig.2. Confusion Matrix for OCR Performance

The confusion matrix illustrates the performance of the OCR module in recognizing characters from scanned PYQ documents. The CNN-based OCR model records 900 true positives and 132 true negatives, showing strong recognition capability. Misclassification errors (100 false negatives and 68 false positives) mainly occur in low-resolution or handwritten segments. These results confirm that the OCR system maintains high reliability with an overall accuracy of 86%, ensuring that extracted text is sufficiently accurate for NLP-based topic identification.

Table 2. Performance Summary Table

Component	**Metric/Performance**
CNN-based OCR	86%
SBERT-based Clustering	73%
Processing Time	40 seconds

Table 2 summarizes the performance metrics of the major modules in the *Learnova* pipeline. The OCR component performs with high accuracy, the keyword extraction module achieves high relevance, and topic ranking matches expert identification with high consistency. Overall, the results verify the robustness and effectiveness of the system in delivering accurate, exam-oriented insights.

Table 3. Comparative Performance Analysis with Existing PYQ Analysis Approaches

System	**OCR Used**	**Topic Identification**	**Semantic Clustering**	**Accuracy**	**Processing Time**
Manual PYQ Analysis[5]	None	None	None	~60%-80%(But high fatigue)	Several hours
OCR + Keyword Search[1][4]	Tesseract	Keyword Matching	None	~70%	~2–3 mins
ML-based QA Systems[1][4][7]	OCR + NLP	Rule-based / ML	Limited	~75–80%	~1–2 mins
Learnova (Proposed)	CNN-OCR	RAKE + TextRank	SBERT	**86%**	**40 sec**

The performance metrics for Manual Analysis and Keyword-based systems (Table 3) are derived from baseline benchmarks and standard student workflow studies identified in existing literature [1, 4, 5], while *Learnova*'s metrics are based on our controlled experimental evaluation. Unlike traditional manual PYQ analysis, which is prone to human error and significant time investment, *Learnova*'s integration of SBERT and CNN-OCR provides a structured, highspeed alternative. While manual analysis and basic keyword-matching systems achieve limited accuracy and require several hours, *Learnova* reduces the cognitive load by automating the process in just 40 seconds with 86% accuracy. The comparison

demonstrates that *Learnova* outperforms existing PYQ analysis approaches in terms of processing time and OCR accuracy while maintaining competitive topic consistency, highlighting its suitability for time-efficient exam preparation.

6 Conclusion

The study presents *Learnova* : Smart PYQ Analyzer, an ML-driven system designed to automate the extraction, analysis, and summarization of Previous Years' Question Papers (PYQs). By integrating OCR, NLP-based topic identification, frequency analysis, and transformer-based models, the system effectively converts unstructured PYQ documents into organized, actionable insights. Experimental results demonstrate that the proposed approach achieves high accuracy in text extraction, strong topic relevance detection and high-quality summarization suitable for exam preparation. The system successfully identifies high-weightage topics with a consistency rate of around 73% when compared with manual expert analysis. Overall, *Learnova* significantly reduces the time and effort required for manual PYQ evaluation and enhances students' ability to prepare strategically for examinations. Furthermore, the comparative results confirm that the proposed ML-driven architecture offers a 90% faster alternative to traditional methods while maintaining superior topic consistency, demonstrating its effectiveness and potential scalability for exam preparation than current manual or basic OCR systems.

7 Future Scope

While *Learnova* achieves strong performance in automating PYQ analysis, several enhancements can further improve its capability. The system can be extended to support multi-subject and multi-language question papers using advanced OCR and multilingual NLP models. Incorporating deep-learning-based topic classification can improve the accuracy of topic grouping and weightage prediction. Future versions can integrate student performance analytics to generate personalized study recommendations based on frequently missed or challenging topics. Additionally, deploying the system as a mobile application or cloudbased service would improve accessibility and allow real-time PYQ processing. Further improvements in summarization using larger generative AI models may provide even more refined and context-aware explanations. With these enhancements, *Learnova* has the potential to become a fully intelligent, adaptive learning assistant for students and educational institutions.

References

1. Sowmiya, R.K., Sridevi, S., Sageengrana, S.: Question paper analysis using machinelearning. International Journal of Novel Research and Development (2024)

2. Rodr´ıguez-Ortiz, M.A., Santana-Mancilla, P.C., Anido-Rifo´n, L.E.: Machine Learn-´ ing and Generative AI in Learning Analytics: A Systematic Review. Applied Sciences 15(431) (2025)
3. Liu, M.: Study on Exam Paper Generation Methods and Experiments based on Machine Learning. IECT (2024)
4. Chakankar, T., Shinkar, T., Waghdhare, S., Waichal, S., Phadtare, M.M.: Automated Question Generator Using NLP. International Journal for Research in Applied Science and Engineering Technology (2023)
5. Mulla, N., Gharpure, P.: Automatic question generation: a review. Progress in Artificial Intelligence (2023)
6. Chavan, A., Lone, A., Shejwal, T., Chordiya, N., Gaikwad, M.: Automatic Question Paper Generation Using Bloom's Taxonomy. International Journal of Innovative Research in Management, Pharmacy and Sciences (2023)
7. Tesseract OCR Engine: Tesseract Open Source OCR. Google Developers (2023)
8. IIT Guwahati: GATE 2017-2024 Question Papers and Answer Keys. GATE 2026 Official Website (2026)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of DeepBidirectional Transformers for Language Understanding. NAACL-HLT (2019)
10. Sharma, P., Kaur, M.: Automated assessment systems using NLP and machinelearning: A survey. Journal of Educational Computing Research (2024)
11. Singh, R., Patil, S.: OCR-based educational document analysis using deep learning.Multimedia Tools and Applications (2024)
12. Zhang, Y., Liu, H.: Deep learning approaches for exam question difficulty estimation. IEEE Transactions on Learning Technologies (2024)
13. Li, J., Sun, X.: Automatic exam paper construction using reinforcement learningand NLP. Knowledge-Based Systems (2024)