International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Recent Advances in Computer Applications and Information Technology
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

AI-Based Syllabus Analysis and Topic Classification System

Shantanu Raut, Vansh Dhodare

G H Raisoni University, Amravati, Maharashtra, India

Abstract

The rapid expansion of academic programs, interdisciplinary courses, and outcome-based education frameworks across universities and educational boards has significantly increased the complexity of syllabus management and curriculum analysis. Institutions are required to continuously update course content to align with industry standards, accreditation requirements, and evolving technological trends. However, traditional manual methods of reviewing and categorizing syllabus documents are time-consuming, inconsistent, and prone to human error. To address these challenges, this research proposes an AI-Based Syllabus Analysis and Topic Classification System that automates the process of analysing, organizing, and classifying syllabus content using advanced Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques The proposed system is designed to extract textual information from digital syllabus documents in formats such as PDF and DOCX and convert them into machine-readable structured data. The extracted text undergoes multiple preprocessing stages, including tokenization, stop-word removal, stemming, and lemmatization, to enhance data quality and reduce noise. Feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) [9] are employed to convert textual data into numerical vectors that represent the importance of terms within the syllabus corpus. These feature vectors are then used to train and evaluate various machine learning classification algorithms The system incorporates supervised learning models such as Naïve Bayes [15], Support Vector Machines (SVM), and selected Deep Learning [5] architectures to categorize syllabus content into predefined academic domains such as Artificial Intelligence, Data Science, Computer Networks, Software Engineering, and others. Comparative performance analysis is conducted to determine the most efficient model in terms of accuracy, precision, recall, and F1-score. Experimental results indicate that machine learning–based approaches significantly improve classification accuracy while efficiently handling large-scale syllabus datasets. Deep learning models, in particular, demonstrate strong performance in capturing contextual relationships among topics in addition to topic classification, the system provides similarity analysis between syllabi from different institutions. This feature enables curriculum comparison, identification of content gaps, detection of redundancies, and benchmarking against standardized frameworks. The automated analysis supports curriculum designers and academic administrators in maintaining uniformity, ensuring compliance with accreditation bodies, and facilitating outcome-based education (OBE) planning. By reducing manual workload and enhancing consistency, the system contributes to improved academic governance and data-driven decision-making.

KEYWORDS*: Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML), Text Classification, Topic Modeling, Educational Data Mining, Curriculum Analysis, Syllabus Classification.*

1. Introduction

In the modern education system, the volume and complexity of academic syllabi have increased significantly across schools, colleges, and universities. With the rapid advancement of knowledge in various domains, curricula are continuously updated to include new topics, interdisciplinary concepts, and skill-based learning outcomes. However, managing, analysing, and organizing syllabus content manually has become a challenging and time-consuming task for educators, academic institutions, and curriculum designers. Traditional methods of syllabus analysis often rely on manual review, which can be prone to inconsistencies, subjective interpretations, and inefficiencies. To address these challenges, Artificial Intelligence (AI) [18] offers innovative and automated solutions.

Artificial Intelligence has emerged as a transformative technology across multiple sectors, including healthcare, finance, transportation, and education. In the educational domain, AI enables intelligent systems capable of understanding, processing, and analysing textual information at scale. One of the promising applications of AI in education is automated syllabus analysis and topic classification. An AI-Based Syllabus Analysis and Topic Classification System leverage Natural Language Processing (NLP) [4], Machine Learning (ML)[1], and data mining [3] techniques to extract, interpret, categorize, and organize syllabus content efficiently. Syllabi serve as foundational documents that outline course objectives, learning outcomes, topics, subtopics, assessment methods, and reference materials. They provide a structured framework for both instructors and students. However, when institutions manage hundreds or thousands of courses across different departments, comparing syllabi, identifying overlaps, ensuring standardization, and updating content becomes increasingly complex. Manual analysis not only consumes valuable time but may also result in redundancy, gaps in learning outcomes, or misalignment with industry requirements. Therefore, an intelligent automated system is essential for effective syllabus management.

The proposed AI-Based Syllabus Analysis and Topic Classification System aim to automate the process of analysing syllabus documents and classifying topics into predefined or dynamically generated categories. Using Natural Language Processing techniques such as tokenization, stop-word removal, stemming, and lemmatization, the system processes textual data from syllabus documents. Advanced machine learning algorithms such as Naïve Bayes [15], Support Vector Machines (SVM)

[16], Decision Trees, or deep learning models like Neural Networks [14] and Transformers can be used to classify topics accurately into subject domains, difficulty levels, or competency categories. One of the major advantages of this system is its ability to identify relationships between topics across different courses. For example, it can detect overlapping concepts between Computer Science, Data Science, and Artificial Intelligence syllabi. This helps institutions reduce duplication and promote interdisciplinary learning. Additionally, the system can highlight missing key concepts by comparing syllabi against standard curriculum frameworks or industry benchmarks. As a result, academic institutions can ensure that their courses remain relevant and up-to-date.

Another important aspect of the system is scalability. Educational institutions often deal with syllabus documents in various formats such as PDF, DOCX, or web-based text. The AI system can integrate document parsing tools to extract textual content from different file formats. Once extracted, the content is structured into meaningful segments like units, modules, or chapters. These segments are then classified into relevant topic categories using trained machine learning models. The system can continuously improve its classification accuracy through model training and feedback mechanisms. Furthermore, the AI-Based Syllabus Analysis and Topic Classification System support data-driven decision-making. Administrators can generate reports and visualizations that show topic distribution across departments, emerging trends in course design, and alignment with academic standards. For example, the system can identify whether a syllabus emphasizes theoretical concepts more than practical skills or whether emerging technologies such as Artificial Intelligence, Blockchain, or Cloud Computing are sufficiently covered. Such insights help curriculum developers make informed improvements.

The integration of AI in syllabus analysis also supports personalized education. By categorizing topics based on difficulty levels or learning outcomes, the system can assist in designing adaptive learning paths for students. For instance, beginner-level topics can be separated from advanced topics, enabling students to follow a structured progression. Additionally, AI-driven topic classification can support recommendation systems that suggest supplementary learning resources based on syllabus content. From a technical perspective, the system architecture typically includes modules for data collection, text preprocessing, feature extraction, classification, and result visualization. Feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), word embeddings [6] (Word2Vec, Glove), or contextual embeddings (BERT) enhance the model's ability to understand semantic relationships within syllabus text. Classification models are trained on labelled datasets to ensure accurate topic categorization. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the system's effectiveness.

Despite its advantages, implementing such a system requires addressing certain challenges. Variations in syllabus writing styles, ambiguity in terminology, and domain-specific jargon may affect classification accuracy. Additionally, maintaining data privacy and ensuring ethical use of AI are critical considerations. Continuous model training and validation are necessary to improve reliability and adaptability to evolving academic standards. In conclusion, the AI-Based Syllabus Analysis and Topic Classification System represent a significant step toward intelligent curriculum management in the education sector. By automating syllabus analysis, reducing manual effort, ensuring content consistency, and providing valuable insights, the system enhances academic planning and quality assurance. The application of AI technologies such as NLP and Machine Learning enables efficient processing of large volumes of syllabus data while supporting innovation in curriculum design. As educational institutions increasingly adopt digital transformation strategies, AI-driven syllabus analysis systems will play a crucial role in shaping future-ready learning environments. The rapid expansion of academic disciplines and the continuous evolution of knowledge have significantly increased the volume and complexity of syllabi across educational institutions. Schools, colleges, and universities frequently revise their curricula to incorporate emerging technologies, interdisciplinary approaches, and industry-oriented skills. As a result, managing and analysing syllabus documents manually has become both time-consuming and inefficient. Traditional review methods often involve subjective interpretation, which can lead to inconsistencies, duplication of topics, and gaps in learning outcomes. To overcome these challenges, Artificial Intelligence (AI) provides an effective and automated solution for syllabus management.

An AI-Based Syllabus Analysis and Topic Classification System utilize advanced technologies such as Natural Language Processing (NLP), Machine Learning (ML), and data mining to process and organize syllabus content systematically. Syllabi typically include course objectives, modules, key topics, assessment strategies, and recommended resources. When institutions handle numerous courses across multiple departments, ensuring alignment, avoiding redundancy, and maintaining academic standards becomes complex. AI-driven systems can automatically extract textual data from various document formats, preprocess the content through tokenization and normalization, and classify topics into predefined or dynamically generated categories using algorithms like Naïve Bayes, Support Vector Machines, or deep learning [5] models. Such a system enhances curriculum planning by identifying overlapping concepts across subjects, detecting missing competencies, and aligning courses with industry standards. It also supports data-driven decision-making by generating analytical reports on topic distribution and skill emphasis. Moreover, by categorizing topics based on difficulty levels and learning outcomes, the system can contribute to personalized and adaptive learning pathways. Overall, AI-powered syllabus analysis improves efficiency, accuracy, and consistency in academic planning while supporting innovation and quality assurance in modern education.

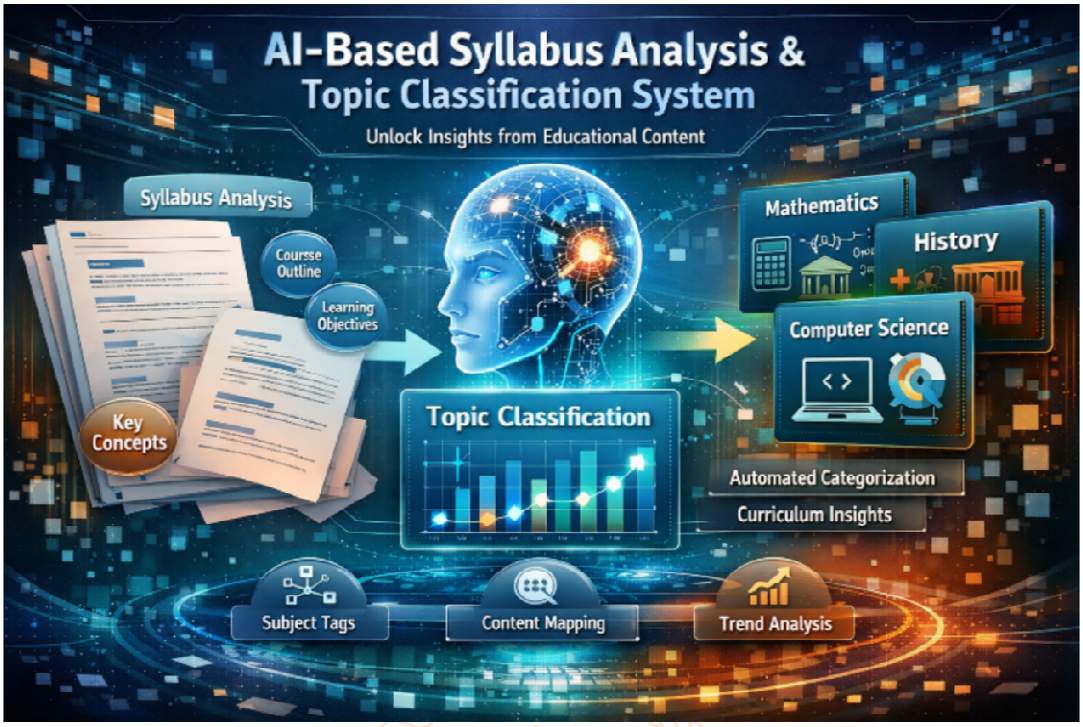

Figure 1. Conceptual Model of AI-Driven Curriculum Analysis System.

2. Literature Review

A. Text Mining and Syllabus Understanding

Text mining techniques [11] have been widely used for educational content analysis, enabling extraction of meaningful information from unstructured text. Studies (e.g., mining syllabi, course descriptions) show that NLP [4] methods like tokenization, stemming, and semantic analysis help identify key concepts and topics from curriculum documents. Earlier work has focused on keyword extraction from educational resources, improving searchability and organization of content. Gap: Limited research specifically targets automatic classification of syllabus topics into structured categories or learning objectives.

B. NLP for Curriculum Analysis

Research has shown that NLP algorithms can automatically process educational texts to identify semantic topics using LDA (Latent Dirichlet Allocation) [8], TF-IDF [9], and word embeddings. Topic Modeling has been applied to cluster related syllabus topics, enabling curriculum mapping and identifying overlaps. Some studies have utilized linguistic analysis to detect learning outcomes and syllabus structure from academic documents. Gap: Many methods focus on unsupervised topic discovery but lack accurate classification aligned with predefined educational taxonomies.

C. Machine Learning in Educational Document Classification

Machine Learning (ML) classifiers like Naïve Bayes [15], SVM [16], and Random Forest have been successfully used in educational text classification— e.g., classifying academic papers, assignments, and educational forums. Results suggest ML models outperform manual classification, offering scalability and consistency. Gap: Limited studies employ deep learning-based classification for syllabus content due to data availability constraints.

D. Deep Learning and Semantic Understanding

Recent advancements such as neural networks and transformer models (e.g., BERT) [7] have significantly improved semantic comprehension of complex educational text. These approaches capture context and relational meaning, which enhances topic classification accuracy. Gap: Though promising, few implementations exist that customize deep learning [5] specifically for syllabus topic categorization.

E. Curriculum Mapping and Alignment

Prior research on curriculum alignment emphasizes comparing course outcomes with educational standards using rule-based and statistical methods. Automated curriculum mapping tools exist but often require manual tagging and human supervision.

3. Research Methodology

A. Research Design

The proposed AI-Based Syllabus Analysis and Topic Classification System follow a structured system development methodology that integrates Natural Language Processing (NLP) and Machine Learning (ML) techniques. The research design is both experimental and developmental in nature, as it involves designing a functional system, training classification models, and evaluating their performance using real academic data. The primary objective of the system is to automatically analyse syllabus documents and classify topics into predefined academic categories with high accuracy and reliability. The research workflow is divided into several systematic phases: data collection, text extraction, preprocessing, feature engineering, model development, classification, and evaluation. Each stage builds upon the previous one to ensure consistency and performance optimization. In the initial stage, syllabus documents are collected from multiple academic sources. In the next phase, raw text is extracted and transformed into structured machine-readable format. The cleaned and processed data is then converted into numerical representations through feature extraction techniques such as TF-IDF or word embeddings.

Supervised learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Random Forest, or deep learning models like LSTM and transformer-based models can be used for classification. If labelled data is available, the dataset is divided into training and testing sets (e.g., 80% training, 20% testing). The system learns patterns from labelled examples and predicts categories for unseen syllabus content. Model evaluation is performed using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques may also be used to ensure robustness and avoid overfitting. This research design ensures scalability and adaptability, allowing the system to be extended to additional subject domains or updated syllabus formats in the future.

B. Data Collection

Data collection is the foundational stage of the AI-Based Syllabus Analysis and Topic Classification System. The quality, diversity, and representativeness of the dataset significantly influence the performance of the classification model. In this phase, syllabus documents are gathered from universities, colleges, online academic repositories, and official educational websites. Documents may exist in multiple digital formats such as PDF, DOCX, TXT, or HTML The collected syllabi typically contain structured academic information including course titles, course objectives, learning outcomes, unit-wise topics, references, credit distribution, and evaluation patterns. To ensure diversity, syllabi are collected from various domains such as Computer Science, Commerce, Management, Engineering, Arts, and Science. This multi-domain dataset helps the model learn generalized patterns rather than being biased toward a single subject area.

Since documents may be available in different formats, text extraction tools are used to convert them into machine-readable format. For PDF and DOCX files, libraries such as PyPDF2 or python-docx can be used to extract text. For HTML documents, web scraping tools and parsers such as Beautiful Soup may be utilized. In cases where documents are scanned images, Optical Character Recognition (OCR) techniques such as Tesseract OCR are applied to convert images into editable text After extraction, the text is organized into a structured database. Each document is stored along with metadata such as institution name, subject domain, academic year, and course code. If a supervised learning approach is adopted, a portion of the dataset is manually labelled into predefined categories such as Programming, Database Systems, Artificial Intelligence, Networking, Software Engineering, or Data Science. This labelled dataset acts as ground truth for model training and validation A carefully designed data collection strategy ensures that the dataset is comprehensive, balanced across categories, and suitable for building an accurate and reliable classification system.

C. Data Preprocessing

Data preprocessing is a crucial phase in transforming raw syllabus text into structured and meaningful data suitable for machine learning algorithms. Raw academic documents often contain noise, formatting inconsistencies, irrelevant symbols, and redundant information. Without proper preprocessing, these irregularities can negatively impact model performance.

The preprocessing pipeline begins with text cleaning. Special characters, extra spaces, HTML tags, headers, footers, and irrelevant formatting symbols are removed. All text is converted to lowercase to maintain uniformity and prevent duplication of words due to case sensitivity. Numerical values that do not contribute to semantic meaning may be removed unless they represent meaningful academic content.

Next, tokenization is performed. Tokenization involves breaking down the text into smaller units such as words or sentences. Word-level tokenization is commonly used for text classification tasks. After tokenization, stop-word removal is applied. Stop words are commonly occurring words such as "the," "is," "and," and "of," which do not contribute significant meaning to the classification process. Removing stop words reduces dimensionality and improves computational efficiency.

Stemming and lemmatization are important preprocessing techniques used to normalize words. Stemming uses rule-based truncation methods to remove suffixes and reduce words to their base forms. For example, "computing," "computer," and "computed" may be reduced to "compute." However, stemming may sometimes produce non-dictionary words. Lemmatization, on the other hand, uses vocabulary knowledge and morphological analysis to return meaningful root words. For example, "studies" becomes "study," and "running" becomes "run." Lemmatization generally provides more accurate linguistic normalization compared to stemming.

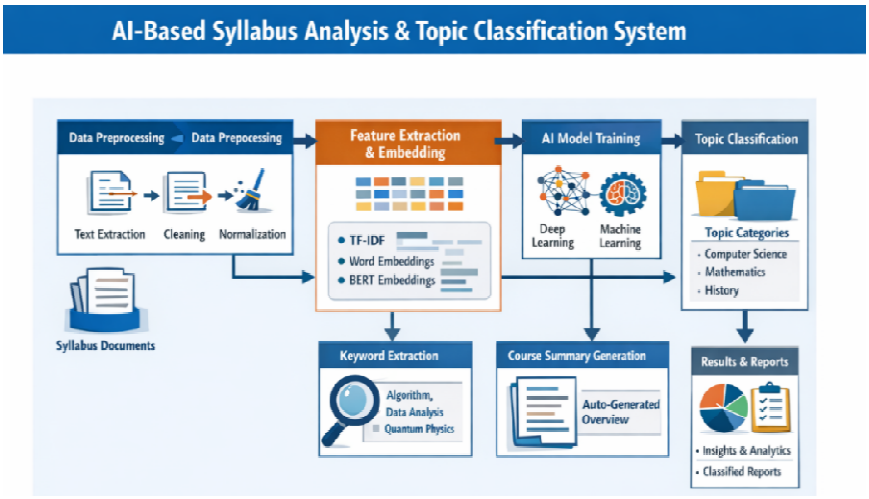

Figure 2. Block Diagram of AI-Based Curriculum Analysis System

4. Result

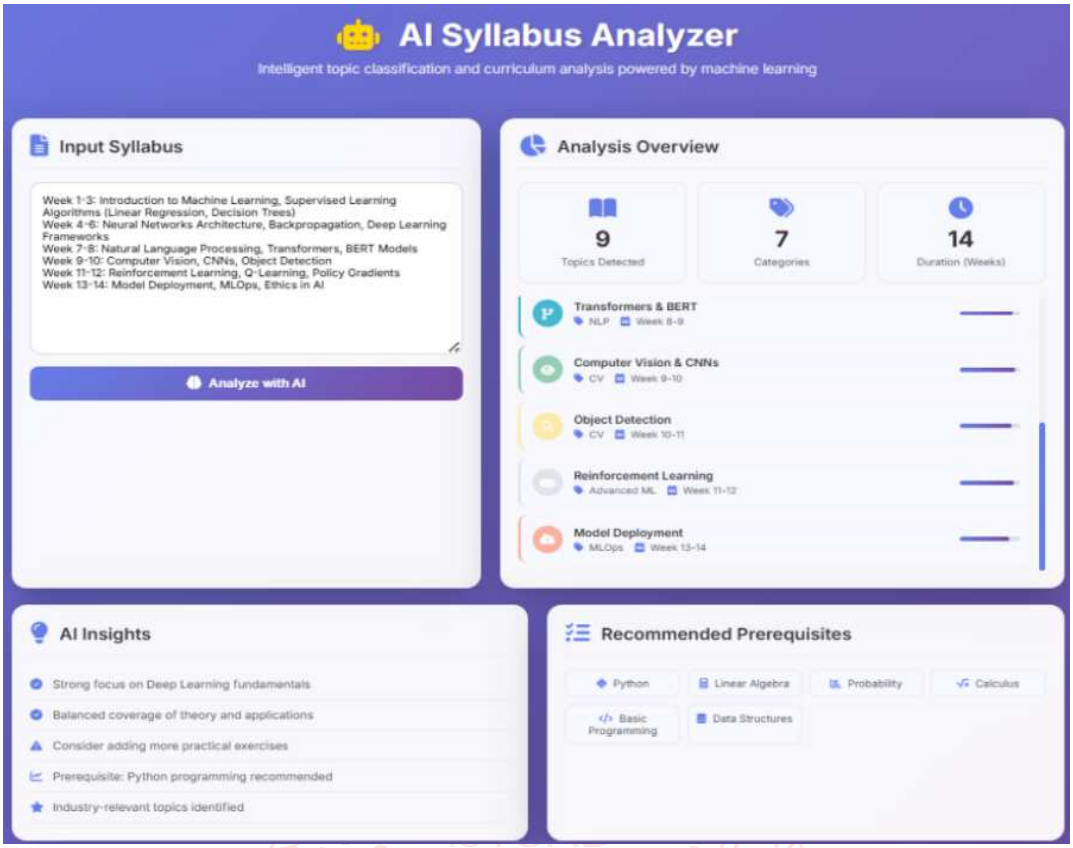

Figure 3. Automated Syllabus Topic Extraction and Learning Insights System

5. Conclusion

The AI-Based Syllabus Analysis and Topic Classification System represent a transformative advancement in the integration of Artificial Intelligence within the modern education ecosystem. As educational institutions strive to remain competitive and globally relevant, they are continuously revising curricula to incorporate emerging technologies, interdisciplinary approaches, competency-based learning models, and industry-aligned skill development. This rapid evolution has led to a substantial increase in the volume, diversity, and complexity of syllabus documents across departments and programs. Managing such extensive academic content through traditional manual methods has become inefficient, error-prone, and unsustainable. Manual review processes often lack standardization, consume significant administrative time, and make it difficult to identify redundancies, gaps, or inconsistencies across courses. In this context, the proposed AI-driven system offers an intelligent, automated, and scalable solution that modernizes curriculum management and enhances institutional effectiveness.

The system leverages advanced Natural Language Processing (NLP) [4] and Machine Learning (ML) techniques to process large volumes of unstructured syllabus text efficiently. Syllabus documents typically include course descriptions, objectives, learning outcomes, topic outlines, assessment methods, reference materials, and skill competencies. However, since this information is usually presented in free-text format, extracting structured insights manually is highly challenging. Through systematic preprocessing steps such as tokenization, stop-word removal, normalization, stemming, and lemmatization, the system cleans and standardizes textual data to ensure consistency and reduce noise. These steps are crucial in preparing the dataset for meaningful computational analysis and improving model performance.

Feature extraction plays a crucial role in improving the intelligence and accuracy of the proposed syllabus analysis system. Techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) and Bag-of-Words convert textual syllabus data into structured numerical representations, enabling efficient computational processing. More advanced approaches, including Word2Vec and contextual embedding models, further enhance analytical depth by capturing semantic meaning and contextual relationships between words. These methods allow the system to recognize conceptual similarities even when different terminology is used. For instance, terms like “machine learning,” “predictive Modeling,” and “data-driven algorithms” can be identified as closely related within the same academic domain.

After extracting meaningful features, classification algorithms such as Naïve Bayes, Support Vector Machines, Random Forests, and deep learning models categorize topics into predefined subject domains or competency levels. This structured classification improves curriculum transparency and quality assurance by identifying redundant modules, overlapping concepts, and missing foundational topics. The system also facilitates alignment with accreditation standards and national education frameworks.

Additionally, automated processing of diverse document formats and generation of visual analytical dashboards enable data-driven decision-making. By reducing manual effort and highlighting curriculum trends and gaps, the system supports strategic academic planning and interdisciplinary collaboration across institutions.

Reference

[1] T. M. Mitchell, "Machine Learning, 1997," McGraw-Hill, New York, NY, USA.

[2] C. M. Bishop, "Pattern Recognition and Machine Learning, 2006," Springer, New York, NY, USA.

[3] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques, 2011," 3rd ed Morgan 2011.

[4] D. Jura sky and J. H. Martin, "Speech and Language Processing, 2009," 2nd ed., Pearson.

[5] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning, 2016," MIT Press.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019," in Proc. NAACL-HLT.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation, 2003," Journal of Machine Learning Research, vol. 3, pp. 993–1022.

[9] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval, 1988 Information Processing & Management, vol. 24, no. 5, pp. 513–523.

[10] F. Sebastiani, "Machine Learning in Automated Text Categorization, 2002," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47.

[11] R. Feldman and J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data, 2007," Cambridge University Press.

[12] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art, 2010," IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601–618.

[13] R. S. Baker and K. Yake, "The State of Educational Data Mining in 2009: A Review and Future Visions, 2009," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17.

[14] S. K. Pal and S. Mitra, "Multilayer Perceptron, Fuzzy Sets, and Classification, 1992," IEEE Transactions on Neural Networks, vol. 3, no. 5, pp. 683–697.

[15] A. McCallum and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification, 1998," AAAI Workshop on Learning for Text Categorization.

[16] C. Cortes and V. Vanik, "Support-Vector Networks, 1995," Machine Learning, vol. 20, pp. 273–297.

[17] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1998," in Proc. European Conference on Machine Learning (ECML).

[18] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach, 2010," 3rd ed., Pearson.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning, 2015," Nature, vol. 521, no. 7553, pp. 436–444.

[20] A. Vaswani et al., "Attention Is All You Need, 2017," in Proc. Advances in Neural Information