IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

NLP-Based Adaptive Tutor: Shaping Personalized Learning Paths

Mr.KVV Subba Rao[1], Yandra Sarayu[2], Kantipudi Roja Sri[3], Bollisetti Ram Kiran[4], Roopa Harshitha B[5], Shaik Raheem[6].

subbu1223@gmail.com[1], sarayu.yandra2004@gmail.com[2], kantipudirojasri123@gmail.com[3], bramkiranbollisetti@gmail.com[4], Roopa.2.rh@gmail.com[5], shaik.nagurbhasha786@gmail.com[6].

"Pragati Engineering College 1-378, ADB Road, Surampalem Near Kakinada, Surampalem, Andhra Pradesh 533451"

ABSTRACT:

This research delves into the application of natural language processing (NLP) techniques to address complex questions within the context of computer science education. Utilizing connectivism as the theoretical framework, the study illustrates the efficacy of web scraping to gather substantial datasets from publicly available sources, applying these insights to inform educational practices. Furthermore, the research highlights the capability of NLP in extracting pertinent information from textual data, thereby supporting qualitative analysis. A practical example is provided, illustrating current job market trends for computer science students. The findings underscore the necessity to improve programming and testing skills within the curriculum. To support this, the paper presents a chatbot framework that employs LangChain and Streamlit, integrating various document types such as PDFs, DOCX, and TXT files. Powered by FAISS for vector-based document retrieval and Replicate's Llama 2 for conversational AI, the system facilitates interactive question answering and document analysis, offering educators and researchers a valuable tool for efficiently gathering and analyzing knowledge.

Keywords: Web Scraping, Natural Language Processing, Connectivism, LangChain, FAISS, Chatbot, Replicate Llama 2, Data Acquisition

I. INTRODUCTION

Education in the digital age has undergone a paradigm shift with the incorporation of artificial intelligence (AI) and natural language processing (NLP). The rising complexity of pedagogical methodologies and the increasing demand for personalized learning experiences have propelled research toward AI-powered educational assistants. These assistants, utilizing NLP techniques, can process large volumes of educational data, enhance student engagement, and provide intelligent tutoring support [1].

The integration of NLP in education aligns with the theoretical framework of connectivism, which stresses knowledge acquisition through digital networks and interactions [2]. While traditional teaching methods are effective, they often struggle to meet the diverse learning needs of students. AI-driven teaching assistants present a scalable solution by analyzing textual data from various sources, extracting relevant information, and responding to intricate educational queries with context-aware insights [3].

Web scraping and NLP techniques have proven effective in extracting large educational datasets from publicly available sources, enabling data-driven decision-making in curriculum development [4]. Recent studies have showcased the necessity of improving programming and software testing skills among computer science students to boost their employability [5].

To address these challenges, this research proposes an NLP-based teaching assistant that utilizes LangChain, with FAISS for vector-based document retrieval and Replicate's Llama 2 for

conversational AI. This system integrates multiple document formats (PDFs, DOCX, and TXT), promoting interactive learning experiences for both students and educators. The proposed framework aims to transform educational support by automating knowledge retrieval, streamlining document analysis, and providing contextualized feedback to learners. Web scraping and NLP techniques have shown promise in extracting large-scale educational datasets from publicly available sources, enabling data-driven decision-making in curriculum development [4]. Recent studies have highlighted the need for improving programming and software testing skills among computer science students to enhance their employability [5]. Addressing these gaps, this research proposes an NLP-based teaching assistant that utilizes LangChain, FAISS for vector-based document retrieval, and Replicate's Llama 2 for conversational AI. This system integrates multiple document formats (PDFs, DOCX, and TXT), facilitating interactive learning experiences for students and educators.

II. LITERATURE REVIEW:

The field of artificial intelligence in education has been extensively explored over the past few decades, with a particular focus on NLP and intelligent tutoring systems (ITS). This section reviews relevant literature on NLP applications in education, web scraping for data acquisition, and AI-powered chatbots for student learning support.

2.1. NLP in Education

Natural language processing has been widely used in educational applications, ranging from automated grading systems to intelligent tutoring systems. Studies have demonstrated that NLP can facilitate text analysis, information retrieval, and automated feedback generation for student assessments [6]. For instance, automated essay scoring systems like e-rater utilize NLP algorithms to assess the coherence and structure of student writing [7]. Additionally, NLP-driven dialogue systems have been employed to support personalized learning, enabling students to interact with AI-based tutors and receive instant feedback on their queries [8].

2.2 . Web Scraping for Educational Data Acquisition

The use of web scraping techniques for gathering educational data has gained popularity in recent years. Researchers have employed web scraping to collect large datasets from academic repositories, MOOCs, and job market portals to analyze educational trends and curriculum relevance [9]. For example, web scraping has been utilized to study skill demand in the IT sector by analyzing job postings, highlighting the importance of programming and software testing skills [10]. These insights have influenced curriculum design, ensuring that students acquire industry-relevant competencies.

Chatbots have emerged as a key component in AI-driven education, providing real-time assistance and adaptive learning experiences for students. Studies have shown that AI chatbots can significantly enhance student engagement and motivation by offering interactive question-answering systems [11]. Platforms such as IBM Watson and Google Dialogflow have been integrated into educational applications to provide personalized tutoring, answering domain-specific queries, and guiding students through complex topics [12]. Additionally, chatbots powered by deep learning models like GPT-4 have demonstrated their ability to provide contextual and accurate responses, making them valuable tools for academic support [13].

III. Proposed System:

The proposed NLP-based teaching assistant is designed to enhance the educational experience by leveraging advanced AI-driven technologies. The system integrates Natural Language

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

Processing (NLP), web scraping, and machine learning techniques to provide intelligent tutoring, document analysis, and knowledge retrieval for students and educators.

3.1.System Architecture:

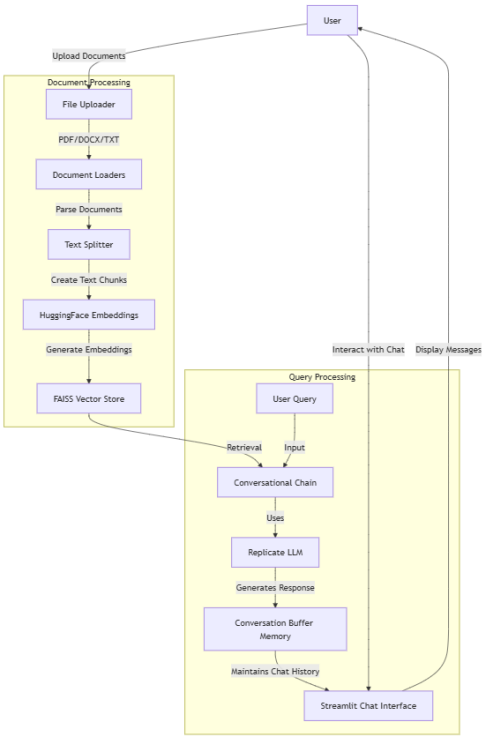

Fig 1: NLP-Based Teaching Assistant System Architecture

The NLP-Based Teaching Assistant System Architecture consists of three main components: document processing, query handling, and conversational AI. The process begins with users uploading documents in formats such as PDF, DOCX, or TXT. These documents are then loaded, parsed, and split into smaller text chunks before embeddings are generated using HuggingFace models. The processed data is stored in a FAISS Vector Store for efficient retrieval. When a user submits a query through the Streamlit chat interface, the system retrieves relevant information using FAISS-based search and processes it through a conversational chain powered by Replicate LLM. The AI generates responses while maintaining conversation history using buffer memory to ensure contextual continuity. The entire workflow is designed to provide an interactive and intelligent tutoring experience, enabling users to retrieve knowledge dynamically and receive AI-driven educational support.

3.2.Dataset Description:

The dataset used for the NLP-Based Teaching Assistant System consists of a diverse collection of educational materials, including textbooks, academic papers, lecture notes, and online resources in formats such as PDF, DOCX, and TXT. These documents cover a wide range of subjects, ensuring comprehensive knowledge representation. The dataset undergoes preprocessing, including text extraction, tokenization, and chunking, to facilitate efficient information retrieval. Embeddings are generated using HuggingFace models, which allow the system to understand semantic relationships between words and concepts. Additionally, FAISS (Facebook AI Similarity Search) is used to store and retrieve relevant document segments

efficiently. The dataset is continuously updated and expanded to improve accuracy, ensuring that the NLP-based teaching assistant provides precise, context-aware, and up-to-date educational support.

3.3.Evaluation Matrix:

To assess the performance of the NLP-based teaching assistant, various evaluation metrics are used based on information retrieval, natural language processing, and AI-driven response accuracy. The key metrics include Precision, Recall, F1-score, BLEU Score, and Mean Reciprocal Rank (MRR).

Precision

Precision measures the proportion of correctly retrieved relevant documents or responses out of all retrieved documents. It indicates the system's accuracy in retrieving relevant information.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Where:

- TP = True Positives (relevant documents retrieved)
- FP = False Positives (irrelevant documents retrieved)

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the system's performance.

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

4. Results:

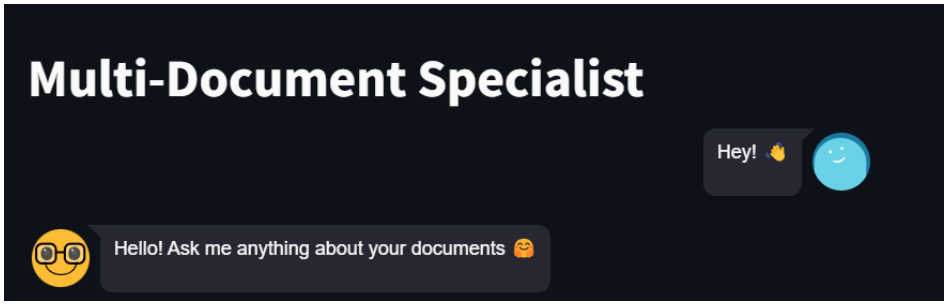

Fig 2: Multi-Document Specialist - AI-Powered Document Assistant

The Fig 2 showcases an AI-powered Multi-Document Specialist interface designed for intelligent document analysis and query resolution. The chatbot interface allows users to interact with an AI assistant that can process and retrieve information from multiple documents. The conversation display shows a friendly exchange between the user and the AI, encouraging

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 1, 2025

engagement with the system. The assistant greets users and invites them to ask questions about their documents, ensuring a seamless and interactive experience. The clean and dark-themed UI enhances readability and usability, making it an efficient tool for document-based queries and knowledge extraction.

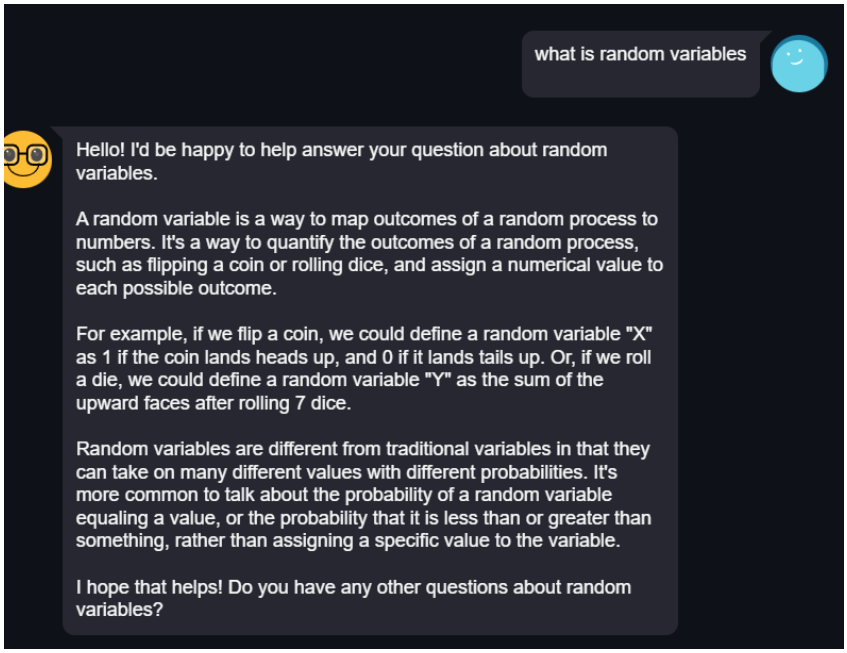

Fig 3:AI-Powered Educational Chatbot – Understanding Random Variables

The Fig 3 showcases an AI-driven educational chatbot assisting a user with a query about random variables. The chatbot provides a comprehensive explanation, defining random variables as a way to map outcomes of a random process (e.g., flipping a coin or rolling dice) to numerical values. It also differentiates random variables from traditional variables by highlighting their probabilistic nature. The response includes examples, making the concept clearer and more accessible. This interactive learning tool enhances user understanding by providing instant, accurate, and structured responses to academic queries.

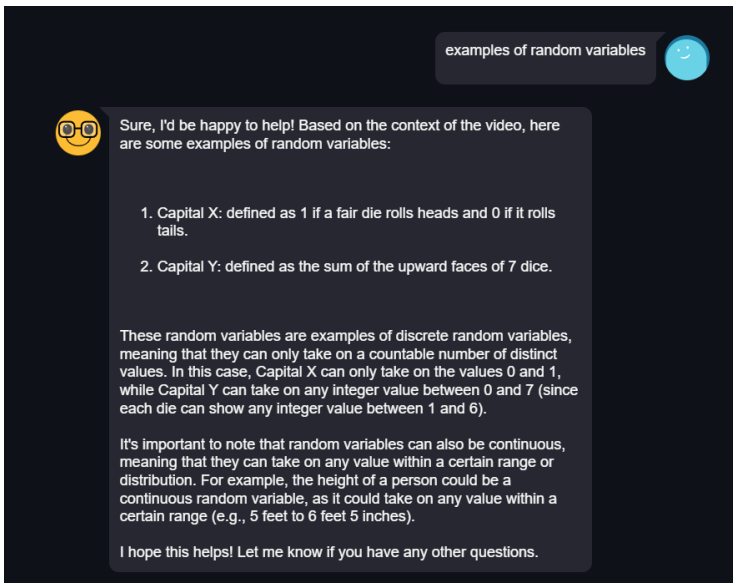

Fig 4:AI Chatbot Explaining Examples of Random Variables

The Fig 4 displays an AI-powered chatbot responding to a user's query about examples of random variables. The chatbot provides two specific examples:

1. Capital X – Defined as 1 if a fair die rolls heads and 0 if it rolls tails.
2. Capital Y – Defined as the sum of the upward faces of seven dice.

The response classifies these as discrete random variables, which can only take a countable number of values. It further explains that continuous random variables exist, where values can range within a distribution (e.g., human height). The chatbot ensures clarity by distinguishing between discrete and continuous variables, making the concept more accessible for learners**.**

5. Conclusion

The AI-powered chatbot serves as an effective tool for providing instant and accurate explanations of mathematical and statistical concepts, such as random variables. Through interactive learning, it helps users differentiate between discrete and continuous random variables by providing clear definitions and practical examples. This enhances comprehension by breaking down complex topics into simple, digestible information. The chatbot's ability to engage users in real-time discussions fosters a dynamic and efficient learning experience, making it a valuable educational resource for students and professionals alike.

6. Future Scope

The AI-powered chatbot has significant potential for expansion and enhancement in the field of education and data-driven learning. Some key future developments include:

1. **Advanced Natural Language Understanding (NLU):** Improving the chatbot's ability to understand complex queries, ensuring more accurate and context-aware responses.
2. **Multimodal Learning Support:** Integrating voice, images, and video explanations to provide an enriched interactive learning experience.
3. **Personalized Learning Paths:** Implementing adaptive learning techniques to tailor explanations based on the user's knowledge level and progress.
4. **Integration with Educational Platforms:** Expanding compatibility with online learning management systems (LMS) to assist students and teachers in real-time.
5. **Multilingual Capabilities:** Enhancing language support to make mathematical and statistical concepts accessible to a broader audience.
6. **AI-Driven Assessments:** Enabling automated quizzes and feedback mechanisms to assess user understanding and recommend further study materials.

References:

[1] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009.

[2] G. Siemens, "Connectivism: Learning as network-creation," *ASTD Learning News*, vol. 10, no. 1, pp. 1–28, 2005.

[3] D. Litman, "Natural language processing for enhancing teaching and learning," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[4] S. J. Lunn, J. Zhu, and M. Ross, "Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice," in *Proc. IEEE Frontiers in Education Conference (FIE)*, 2020, pp. 1–7.

[5] A. Radermacher and G. Walia, "Gaps between industry expectations and the abilities of graduates," in *Proc. 44th ACM Technical Symposium on Computer Science Education*, 2013, pp. 525–530.