Extracting Significant information using Large Language Models(LLM)

Team Members:

Rakshith Arya - 4NI20CS075

Shashank K R - 4NI20CS094

Shashank Shandilya - 4NI20CS096

Vishal M V - 4NI20CS124

Project Guide:

Dr. Shabana Sultana

SYSTEM ANALYSIS

Existing System

- **Rule-based systems**, dependent on predefined rules, excel in structured domains but falter with the intricacies of natural language, hindering efficiency with diverse and unstructured data due to their rigid nature.

- Traditional machine learning methods such as **SVMs and Random Forests** require manual feature engineering for information extraction, limiting their ability to capture nuanced semantics in natural language, especially with sparse labeled data.

- **Template-based information extraction** relies on predefined templates for specific data extraction. However, these systems are less adaptable to language variations and may face challenges in dynamic environments lacking predefined information structures.

Proposed System

- Our system leverages Large Language Models (LLMs), like OpenAI's GPT series, for information extraction. Unlike traditional methods, LLMs are pre-trained on diverse language data, enabling them to understand and generate human-like language, making them highly effective in extracting relevant information from unstructured text.

- LLMs surpass limitations of predefined rules or manual feature engineering, excelling in handling diverse natural language variations. Their contextual understanding enables versatility across domains, proving valuable in dynamic language structures where rigid template-based systems fall short.

- The proposed system excels by reducing reliance on labeled data, leveraging pre-training on extensive language data for knowledge transfer to new tasks, ensuring adaptability to diverse language structures, and effective performance across applications.

LITERATURE SURVEY

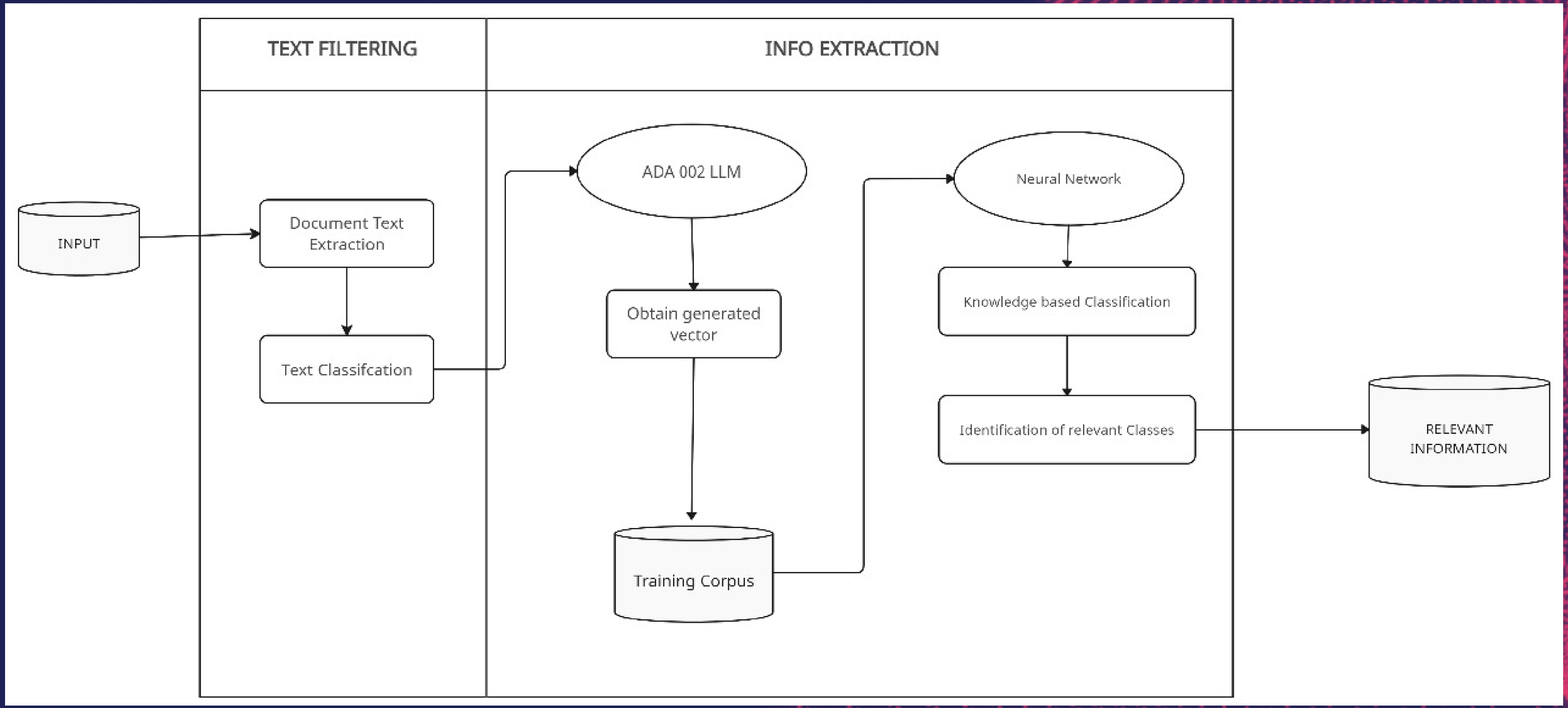
"LMDX: Language Model-based Document Information Extraction and Localization (Perot et al., 2023) introduces a novel approach utilizing Large Language Models (LLMs) for extracting information from visually rich documents (VRDs). The method employs a layout-agnostic prompt and decoding algorithm, enabling precise localization and extraction of various entity types.

Demonstrating state-of-the-art performance and data efficiency, LMDX excels across multiple benchmarks. Another advancement, 'Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning' (Wang et al., 2021), addresses named entity recognition challenges by integrating LLMs with external context retrieval and cooperative learning, enhancing entity identification.

Meanwhile, 'BROS: A layout-aware pre-trained language model for understanding documents' (Hong et al., 2021) focuses on enhancing document understanding through a layout-aware pre-training process, utilizing a novel layout encoding scheme. BROS outperforms layout-agnostic models in various document understanding tasks by effectively capturing spatial relationships between text blocks and content."

SYSTEM DESIGN

Proposed System Architecture



CONCLUSION AND FUTURE WORK

In conclusion, the utilization of Large Language Models (LLMs), exemplified by ADA 002 in this system, has demonstrated profound efficacy in information extraction from legal documents. ADA 002's ability to capture complex semantic meanings through its 1536-dimensional output vector has significantly enhanced the process of classifying, extracting, and understanding textual information. The system's reliance on ADA 002 has proven invaluable in tasks such as text and code search, sentence similarity measurement, and text classification, showcasing its versatility across diverse applications.

Looking forward, the trajectory of LLMs in information extraction holds promising possibilities for further advancements. Future enhancements could involve fine-tuning models on more domain-specific data, potentially improving accuracy in legal document understanding. Additionally, exploring novel pre-training strategies and techniques to enhance the contextual understanding of legal jargon and nuances could lead to more refined and specialized information extraction.


THANK
YOU