IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Intelligent Document Summarisation

[1]Vivek Kumar, [2]Golu Kumar, [3]Md. Salman, [4]Jigar Kumar
[1]B.Tech – CSE, [2]B.Tech – ECE, [3]B.Tech – CSE, [4]B.Tech – CSE
Department of Computer Science and Engineering / Electronics and Communication Engineering
Centurion University of Technology and Management, Paralakhemundi, India

Abstract: This project presents a smart system that reads and understands documents automatically. Users can provide a file in any format — PDF, Word document, image, or plain text — and the system reads it, identifies the content, categorises it, and generates a concise summary. The system also supports a question-answering interface, where users can ask queries and receive responses based on the document content. Additionally, it produces visual analytics such as keyword frequency charts and confidence scores, and allows downloading results as an Excel report. This system is designed for students, researchers, and professionals who handle large volumes of documents, enabling them to extract key information quickly and efficiently.

Index Terms — Document Summarisation, Natural Language Processing, OCR, Machine Learning, Deep Learning, Transformer Models, Text Classification, Keyword Extraction, AI Assistant.

I. INTRODUCTION

Today, we create and store more information than ever before — research papers, reports, articles, scanned files, and much more. The challenge is straightforward: there is too much to read manually. Going through all of it takes a huge amount of time and effort, and is no longer practical in most professional and academic settings.

Recent developments in Artificial Intelligence have made it possible to build systems that can read and understand documents the way a human would — but much faster. These systems can extract key points, identify important words and topics, categorise documents, and even read text from scanned images or photographs of printed pages.

This project builds exactly that kind of system. It combines several AI technologies to create a tool that reads documents, generates clear summaries, identifies the topics covered, highlights the most important keywords, and allows users to ask questions and get direct answers — all through a simple, easy-to-use web interface. No technical knowledge is required; users need only to upload a document.

II. RELATED WORK

Over the past few years, considerable research effort has been directed toward making computers understand documents — not just store them, but actively read and make sense of their content. Early systems used frequency-based methods, analysing how often certain words appeared in a document to infer its topic. These approaches worked at a basic level, but they lacked any genuine understanding of language meaning.

The field advanced with the adoption of machine learning, enabling computers to recognise patterns and classify documents with greater accuracy. However, these techniques still had significant limitations in handling semantic content and context. The real breakthrough arrived with deep learning — specifically, the development of transformer architectures. These models can understand the context of language rather than simply counting words. Contributions from companies such as Google, and platforms such as Hugging Face, made these powerful models widely accessible to researchers and developers worldwide, substantially accelerating progress.

Alongside these advances, complementary technologies were integrated into document processing pipelines — including Optical Character Recognition (OCR), which enables computers to read text from scanned pages and images, and automated keyword extraction, which identifies the most significant terms from a document. However, most of these tools were developed and deployed independently. Researchers typically relied on separate tools for summarisation, classification, and search, with no unified platform bringing them together. This project addresses that gap by integrating OCR, machine learning, deep learning, summarisation, visual dashboards, and an AI assistant into a single coherent platform.

III.RESEARCH METHODOLOGY

The system accepts document input in any common format, including plain text files, PDFs, Word documents, and scanned images. The processing pipeline proceeds through several stages as described below.

3.1 Document Ingestion and Text Extraction:

Upon upload, the system reads the document content. For standard digital files, text is extracted directly from the file structure. For scanned pages or image-based documents, the system applies Optical Character Recognition (OCR) — a technology that visually identifies text and converts it into machine-readable content, analogous to how a human reads a printed page.

3.2 Text Preprocessing

Raw extracted text typically contains considerable noise, including unnecessary symbols, formatting artefacts, and stop words. The system applies standard natural language preprocessing techniques to clean and normalise the text, ensuring it is suitable for downstream analysis.

3.3 Document Classification

Once the text is prepared, the system automatically determines the document category. Classification is performed using a trained machine learning model that has learned to recognise patterns across many document types. This approach yields consistent and accurate categorisation without requiring manual input from the user.

3.4 Summarisation

The system produces two types of summaries. The first is an extractive summary, which selects the most informative sentences directly from the source document. The second is an abstractive summary, which uses advanced AI language models — developed by Google and made available through Hugging Face — to rewrite the document content in natural language. The resulting abstractive summary reads as though composed by a human author.

3.5 Keyword and Topic Extraction

The system identifies the key terms and central themes of each document using Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction. This enables users to quickly understand what a document is about without reading the full content.

3.6 Visual Dashboard and Report Generation

All results are presented through an interactive visual dashboard built with Plotly, displaying charts and graphs that allow users to explore keyword distributions, document similarity measures, and classification confidence scores. A full downloadable report in Excel format is also available, enabling further analysis and offline record-keeping.

IV. RESULTS AND DISCUSSION:

The system was evaluated on a diverse set of documents including text files, PDFs, Word documents, and scanned images, and delivered reliable performance across all document types.

4.1 Text Extraction and Preprocessing

Document reading and text extraction worked reliably for both digitally typed files and scanned inputs. The preprocessing pipeline successfully cleaned the extracted content, making it suitable for subsequent analysis.

4.2 Document Classification

The classification module accurately sorted documents into their correct categories with consistent reliability. The trained model generalised well across diverse document topics and styles.

4.3 Summarisation Performance

Both summarisation approaches delivered useful results. The AI-generated abstractive summaries — powered by Google's transformer models via Hugging Face — produced natural, easy-to-read versions of the original content. The extractive summaries preserved the most important sentences from the source text. Users found both outputs informative and well-formed.

4.4 Keyword and Topic Identification

The keyword and topic extraction module consistently identified the principal themes and important terms from each document, giving users a quick and accurate sense of document content.

4.5 Visual Dashboard

The interactive dashboard provided an effective interface for exploring results. Users could view keyword frequency charts, document similarity comparisons, and classification confidence levels within a single, clean screen. The Excel export functionality was also verified to produce correctly formatted reports for offline use.

4.6 AI Question-Answering

The built-in AI assistant accurately answered user queries about uploaded documents, retrieving relevant information and responding in a natural, conversational manner. The responses were consistently grounded in document content rather than general knowledge.

V. CONCLUSION

This paper presented an intelligent document summarisation system capable of reading, understanding, and transforming documents into useful outputs — including clean summaries, keyword lists, category labels, and direct answers to user queries. The system operates automatically, requiring no technical expertise from the end user.

The platform supports all common file types, including scanned images, and handles the varied and unstructured nature of real-world documents without difficulty. The AI models underlying the system are trained to understand language contextually, which is reflected in the natural quality of the generated summaries and the accuracy of the classifications.

Results demonstrate that the system is accurate, easy to use, and genuinely time-saving. It is well-suited for students processing research literature, professionals managing large document repositories, and analysts working with diverse written content. Future work will explore multilingual support, domain-specific fine-tuning of language models, and real-time collaborative document analysis features.

VI.ACKNOWLEDGMENT:

The authors sincerely thank their project guide for the invaluable advice, patience, and encouragement provided throughout this work. Gratitude is also extended to the Head of the Department and all faculty members of Centurion University of Technology and Management for their guidance and support. The authors also acknowledge the contribution of friends, classmates, and family members whose encouragement made this work possible.

REFERENCES

[1] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8 (Version 8.0.0)," GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[2] S. Yang, W. Wang, S. Gao, and Z. Deng, "Strawberry ripeness detection based on the YOLOv8 algorithm fused with the LW-Swin Transformer," Computers and Electronics in Agriculture, 2024.

[3] R. Mishra and S. Jain, "YOLOv8 for Object Detection: A Comprehensive Review of Advances and Applications in Agriculture," ResearchGate, 2025.

[4] H. Wang, X. Wang, and L. Sun, "FALW-YOLOv8: A Lightweight Model for Detecting Produce Defects in Real-Time," Electronics, vol. 15, no. 1, p. 209, 2026.

[5] S. Jamil et al., "A Systematic Review of Deep Learning-Based Object Detection in Intelligent Agricultural Systems," CMC-Computers, Materials & Continua, 2025.

[6] MDPI Agronomy, "YOLOv8-Driven Detection of Defects, Diseases, and Maturity in Horticultural Products," 2025.

[7] Ultralytics, "YOLOv8: State-of-the-Art Object Detection and Segmentation Models," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proc. IEEE CVPR, 2016, pp. 779–788.

[9] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020.

[10] Google Research, "Google Colaboratory: Cloud-based Jupyter Notebook Environment," 2023. [Online]. Available: https://colab.research.google.com

[11] Kaggle, "Sliced Fruits and Vegetables Dataset," 2022. [Online]. Available: https://www.kaggle.com

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.