

L&T NEUROHACK



SILVER PROBLEM STATEMENT

Develop an innovative text summarization tool that can efficiently process and summarize handwritten texts, all while operating offline.

Team Name: Sonic Coder

Team Lead: Aditya Chavan | 202201814@vupune.ac.in

TEAM MEMBERS



Aditya Chavan

College: Vishwakarma University
Stream: B.Tech (Computer Science)
Year of Graduation: 2026



Saumya Shah

College: Vishwakarma University
Stream: B.Tech (Computer Science)
Year of Graduation: 2026



Bhavin Baldota

College: Vishwakarma University
Stream: B.Tech (Computer Science)
Year of Graduation: 2025



Sharvari Korde

College: Vishwakarma University
Stream: B.Tech (Computer Science)
Year of Graduation: 2026

Problem Statement

Challenge:

Develop a text summarization tool capable of:

- Processing and summarizing handwritten and digital text with a high level of accuracy.
- Operating entirely offline, which is crucial for privacy-sensitive documents and remote areas lacking stable internet.
- Handling a variety of handwriting styles and formats as well as digital text across different file types.
- Generating concise and contextually accurate summaries, even from complex or large texts.
- Running smoothly on devices with a minimum of 16GB RAM to ensure widespread usability.

Why It Matters:

- **Privacy Preservation:** Documents never leave the device, making it suitable for sensitive, private information.
- **Connectivity-Free Operation:** Ideal for users in low-connectivity environments.
- **Efficiency for Document-Heavy Industries:** Automated, private summarization accelerates productivity in fields like healthcare, law, and finance, which often deal with vast amounts of handwritten and printed documentation.

Solution Overview

Solution Components:

1. **PyTorch Foundation:** Selected for its versatility in deep learning and compatibility with offline deployment, PyTorch serves as the base for model development.
2. **Multi-Model Framework:**
 - **The Handwritten Text Model:** is trained on 16,700 samples encompassing a diverse range of handwriting styles. It leverages a custom Convolutional Neural Network (CNN) to accurately capture the distinct characteristics of handwritten text.
 - **Digital Text Extraction Model:** Trained on Google Fonts to ensure recognition across different typefaces and fonts.
 - **Specialized Models:** Models fine-tuned for PDF, JSON, XML, Word, and Excel formats, enabling accurate text extraction across diverse digital files.
3. **Advanced Preprocessing Techniques:** Processes images to enhance text clarity and reduce OCR errors, ensuring quality input for OCR.
4. **OCR with Tesseract:** Google's OCR technology integrates seamlessly to transform processed images into machine-readable text.
5. **Pipeline Workflow Optimization:** Efficiently organizes model flow from input to summary to ensure high performance across various document formats.

System Architecture

System Workflow:

Input Document: Accepts a range of file types (handwritten, PDFs, XML, JSON, etc.).

File Type Detection: Automatically identifies the document type and routes it to the relevant model.

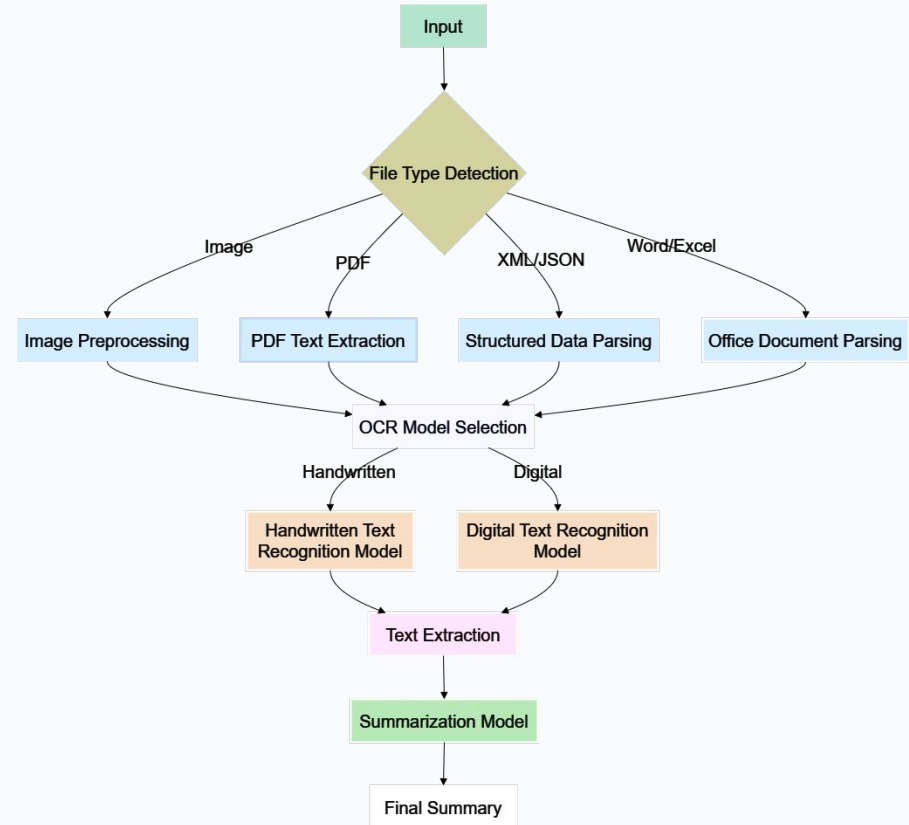
Specialized Models: Assigns the file to the proper model, optimizing extraction accuracy.

Preprocessing: Adjusts image/document quality to enhance OCR performance.
OCR Process: Converts preprocessed data into machine-readable text.

Text Extraction: Retrieves significant text portions while ignoring unnecessary information.

Summarization: Applies NLP algorithms to create a concise, relevant summary.

Final Output: Summarized text is presented in an easily readable format.



Preprocessing Pipeline

BOUNDING BOX DETECTION

Connected Component Analysis: Identifies individual text regions.

Hierarchical Clustering: Groups related text blocks for layout structure.

Multi-Column Handling: Effectively processes text in multi-column or image-wrapped layouts.

RESCALING

Intelligent Resizing: Preserves aspect ratios while resizing.

Benefit: Ensures optimal resolution for OCR without excessive computational load.

BINARIZATION

Adaptive Thresholding: Handles uneven illumination in scanned documents.

Otsu's Method: Selects the optimal threshold for binarization.

Local Binarization: Useful for complex backgrounds to enhance text clarity.

NOISE REMOVAL

Median Filtering: Reduces salt-and-pepper noise.

Morphological Operations: Removes small artifacts around text.

Edge-Preserving Smoothing: Maintains text quality while reducing noise.

DILATION AND EROSION

Customized Kernels: Adjusted for different text types (thin, thick, or broken characters).

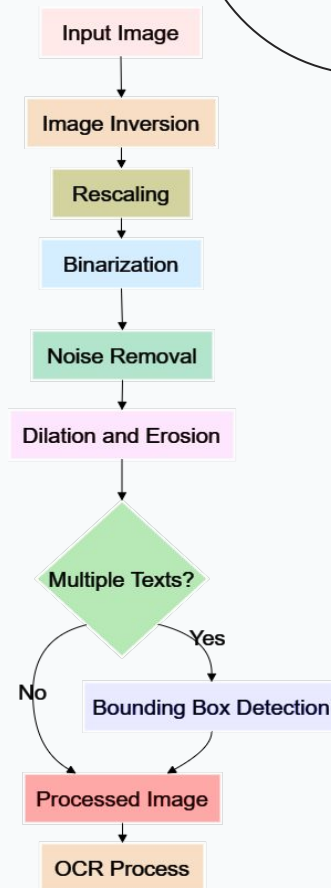
Iterative Application: Closes small gaps in characters.

Controlled Use: Avoids unintentional merging of nearby characters.

INPUT IMAGE

Adaptive Inversion: Based on background-foreground analysis.

Purpose: Improves text visibility, especially effective for dark backgrounds or faded text.



Handwriting Recognition

Custom Handwriting Model Highlights:

Dataset & Training: Trained on 20,000 handwritten samples, covering multiple handwriting styles (cursive, printed, etc.) and various languages.

CNN Architecture: Uses a Convolutional Neural Network (CNN) that specializes in capturing intricate handwriting features.

Data Augmentation: Techniques like rotation, scaling, and distortion help model robustness, making it adaptable to different handwriting styles.

Real-World Testing: Tested on diverse samples to ensure accurate digitization of handwritten content.

Ongoing Updates: Continually updated with additional samples to handle evolving handwriting trends and maintain adaptability.

Digital Text Extraction

Digital Text Recognition Model Details:

Training with Google Fonts: Encompasses over 1,000 fonts in various styles (regular, italic, bold) for high flexibility across text types.

Feature Extraction: Advanced methods to detect subtle differences across fonts, essential for accurate recognition in varied contexts.

Transfer Learning: Uses pre-trained models to enhance accuracy and reduce training time, making it capable of recognizing a wide array of fonts and sizes.

Variation Robustness: Ensures adaptability to different font sizes, colors, and backgrounds.

Regular Model Updates: Incorporates newly released fonts to stay current, boosting the versatility of typed document processing.

OCR Implementation and Summarization

Tesseract OCR Integration

Optimization for offline operation:

- Compressed models for reduced storage and memory footprint.
- Efficient loading and unloading of language-specific data.
- Custom-built dictionaries for improved recognition in specific domains.

Fine-tuning for high-accuracy text extraction:

- Retraining on domain-specific datasets for improved performance.
- Customization of recognition parameters for different document types.
- Integration of post-processing rules for error correction.

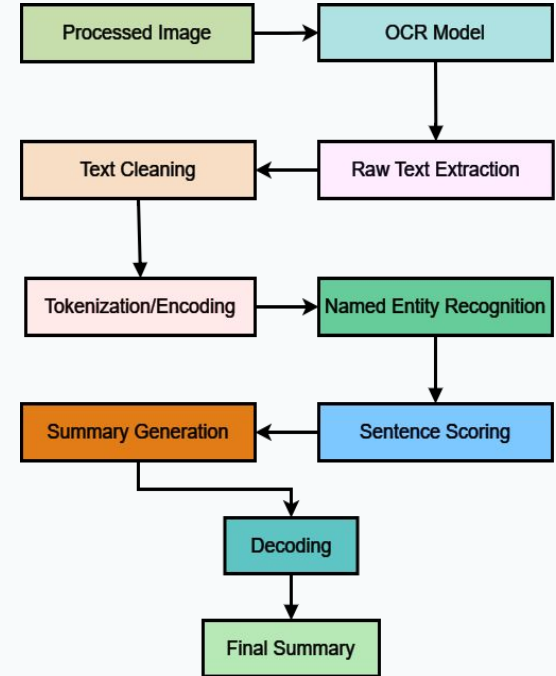
Multi-font support:

- Incorporation of 100+ font packs for diverse document handling.
- Trained on largely available google fonts.

NLP-based Summarization

PyTorch NLP-Based Summarization:

- **Transformer Architecture:** Leverages self-attention for deep contextual understanding, balancing long-term context and detail.
- **Extractive and Abstractive Summarization:** Supports both types to capture the gist or generate new sentences from content.
- **Adaptability to Text Length:** Adjusts summarization density for documents of varying lengths, ensuring clear, coherent summaries.
- **Domain-Specific Tuning:** Fine-tuned to handle specific fields (e.g., legal, healthcare), if needed.
- **Local Processing for Privacy:** Fully offline, so no data is sent externally.
- **Regular Benchmarking:** Evaluated against human-generated summaries for continuous accuracy improvement.



Content Structure Analysis :

- Analyzes the input text's structure, including the number of entities, sentences, tokens, and keywords.
- Determines the summarization style (bullet points, flowchart, table, or paragraph) based on the text's structure and detected elements (e.g., bullet points or arrows).

Optional User Context:

- Optionally prompts the user to provide additional context for better summarization.

Sentiment Analysis :

- Uses Textblob to perform sentiment analysis, returning the polarity (positive/negative/neutral) and subjectivity (objective/subjective) of the text.

Text Cleaning :

- Cleans the input text by decoding any HTML-encoded entities and replacing common HTML tags with readable text.


Cleaning the Summary:

- Post-processes the generated summary to improve formatting.
- Supports different formats based on the chosen style (e.g., bullet points, flowchart).


Section Summarization:

- Summarizes a section of the text using the T5 model.
- Dynamically adjusts the length of the summary based on the length of the input text.
- Uses the cleaned text and optional user-provided context to create an adaptive prompt.

Complete Text Summarization Model

L&T NEUROHACK Text Summarization Tool

Upload Document


Choose a file or drag it here

Supported formats: TXT, PDF, PNG, JPG, JPEG, GIF, DOCX, XLSX, JSON, XML

Context Information (Optional)
Enter any additional context that might help with summarization...

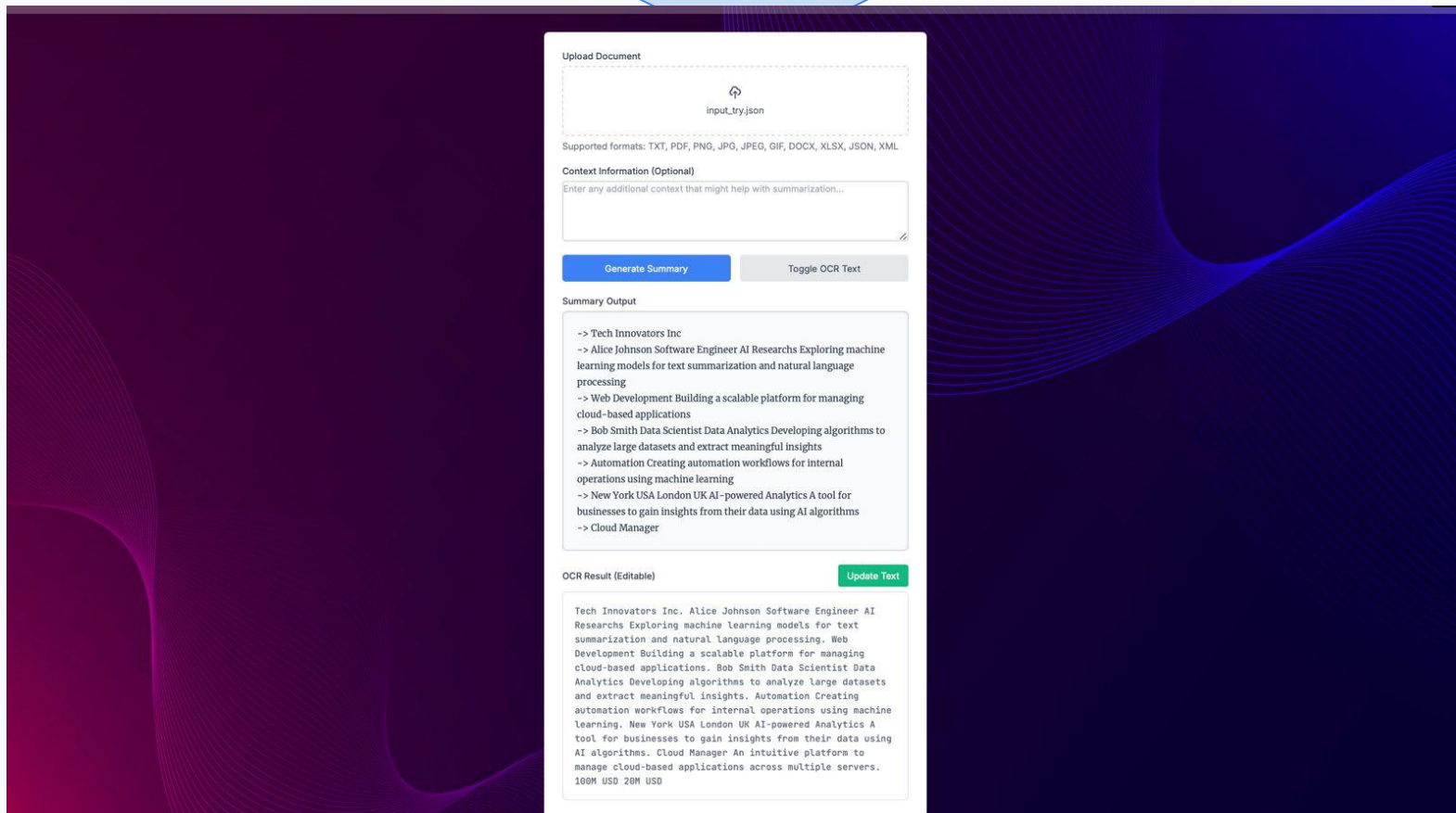
Generate Summary

Toggle OCR Text

Summary Output

Created by:
Aditya Chavan | Bhavin Baldota | Saumya Shah | Sharvari Korde

Complete Text Summarization Model



The interface is a web-based application for text summarization. It features a central white panel on a dark purple background with abstract wavy patterns. At the top of the panel is an 'Upload Document' section with a dashed box containing a file icon and the text 'input_try.json'. Below this, it lists supported formats: TXT, PDF, PNG, JPG, JPEG, GIF, DOCX, XLSX, JSON, XML. The 'Context Information (Optional)' section has a text input field with a placeholder 'Enter any additional context that might help with summarization...'. Two buttons are present: 'Generate Summary' (blue) and 'Toggle OCR Text' (grey). The 'Summary Output' section displays a list of bullet points summarizing the input text. At the bottom, the 'OCR Result (Editable)' section shows the full original text, with an 'Update Text' button to its right.

Upload Document

input_try.json

Supported formats: TXT, PDF, PNG, JPG, JPEG, GIF, DOCX, XLSX, JSON, XML

Context Information (Optional)

Enter any additional context that might help with summarization...

Generate Summary Toggle OCR Text

Summary Output

- > Tech Innovators Inc
- > Alice Johnson Software Engineer AI Researchs Exploring machine learning models for text summarization and natural language processing
- > Web Development Building a scalable platform for managing cloud-based applications
- > Bob Smith Data Scientist Data Analytics Developing algorithms to analyze large datasets and extract meaningful insights
- > Automation Creating automation workflows for internal operations using machine learning
- > New York USA London UK AI-powered Analytics A tool for businesses to gain insights from their data using AI algorithms
- > Cloud Manager

OCR Result (Editable) Update Text

Tech Innovators Inc. Alice Johnson Software Engineer AI Researchs Exploring machine learning models for text summarization and natural language processing. Web Development Building a scalable platform for managing cloud-based applications. Bob Smith Data Scientist Data Analytics Developing algorithms to analyze large datasets and extract meaningful insights. Automation Creating automation workflows for internal operations using machine learning. New York USA London UK AI-powered Analytics A tool for businesses to gain insights from their data using AI algorithms. Cloud Manager An intuitive platform to manage cloud-based applications across multiple servers. 100M USD 20M USD

Complete Text Summarization Model

OCR Result (Editable)

Update Text

Artificial Intelligence and its Applications

Artificial Intelligence (AI) is transforming industries across the globe by introducing advanced capabilities that automate tasks, enhance decision-making, and drive efficiency. From healthcare to finance, AI technologies are being used in a variety of sectors to solve complex problems and deliver innovative solutions.

1. The Evolution of AI

The concept of AI dates back to the mid-20th century, when researchers began exploring the possibility of creating machines that could think and learn. Early developments in AI focused on rule-based systems, where computers followed predefined rules to achieve specific tasks. However, these systems were limited in scope and could only handle simple, structured problems. With the advent of machine learning in the 21st century, AI experienced a significant leap forward. Machine learning allows computers to learn from data without being explicitly programmed, enabling them to handle more complex and unstructured tasks. This breakthrough has led to the widespread adoption of AI across various industries.

2. AI in Healthcare

One of the most promising applications of AI is in the healthcare sector. AI-driven systems are capable of analyzing vast amounts of medical data to assist doctors in diagnosing diseases, predicting patient outcomes, and personalizing treatment plans. For example, AI algorithms can analyze medical images, such as X-rays and MRIs, to detect abnormalities with a high degree of accuracy. AI is also being used to improve drug discovery by predicting how different chemical compounds will interact with the human body. This can significantly speed up the process of developing new treatments for diseases. Additionally, AI-powered virtual assistants are helping patients manage their health by providing reminders to take medication, tracking symptoms, and even answering health-related questions.

3. AI in Finance

The financial industry has embraced AI to automate processes, reduce fraud, and improve decision-making. AI-powered algorithms are used to analyze financial markets and execute trades with minimal human intervention. These systems can process vast amounts of data in real time, enabling

L&T NEUROHACK Text Summarization Tool

Upload Document

input_try.pdf

Supported formats: TXT, PDF, PNG, JPG, JPEG, GIF, DOCX, XLSX, JSON, XML

Context Information (Optional)

Enter any additional context that might help with summarization...

Generate Summary

Toggle OCR Text

Summary Output

Summary:

-> Artificial Intelligence (AI) is transforming industries across the globe by introducing advanced capabilities that automate tasks, enhance decision-making, and drive efficiency

-> From healthcare to finance, AI technologies are being used in a variety of sectors to solve complex problems and deliver innovative solutions

-> From the mid-20th century, researchers began exploring the possibility of creating machines that could think and learn

-> Early developments in AI focused on rule-based systems, where computers followed predefined rules to achieve specific tasks

-> However, these systems were limited in scope and could only handle simple, structured problems

-> With the advent of machine learning in the 21st century, AI experienced a significant leap forward

-> Machine learning allows computers to learn from data without being explicitly programmed, enabling them to handle more complex and unstructured tasks

-> AI-driven systems are capable of analyzing vast amounts of medical data to assist doctors in diagnosing diseases, predicting patient outcomes, and personalizing treatment plans

-> For example, AI-powered virtual assistants are helping patients manage their health by providing reminders to take medication, tracking symptoms, and even answering health-related questions.

OCR Result (Editable)

Update Text