SMART INDIA HACKATHON 2024

- **Problem Statement ID – SIH1678**
- **Problem Statement Title – RE-DACT**
- **Theme – Blockchain & Cyber Security**
- **PS Category – Software**
- **Team ID – 527**
- **Team Name – Team Rookie**

Team Rookie

RE-DACT

SMART INDIA HACKATHON 2024

- **Detailed Explanation:**
  - Any application that blurs out or hides sensitive details such as the name, address etc, eliminating the need to edit the entire document of a file.
  - We made it possible by employing advanced technologies such as Natural Language Processing with Transformer for enhanced accuracy, Optical Character Recognition for handling image-based documents or Microsoft ML to enable TensorFlow and automated redaction for efficient documentation.

- **How it address the problem:**
  - Full concealment of all the identity relating to name, phone number, email address and any other key identifiers.
  - We are creating a software with more flexibility that will be compatible with several file formats such as text, image, pdf, and excel, word which is platform independent can be accessed from android,ios,windows,linux and more

- **Innovation:**
  - Identifying the places of interest by employing spaCy interface and hugging face transformer NER module for named entity recognition.
  - Providing the option of customizable redaction level so the user can select the redaction level that is required by him.
  - Ensuring that school's data are protected and create user friendly interface in order to enhance user experience.
  - Flexible for use with files of other formats such as txt, img, pdf, word, excel.

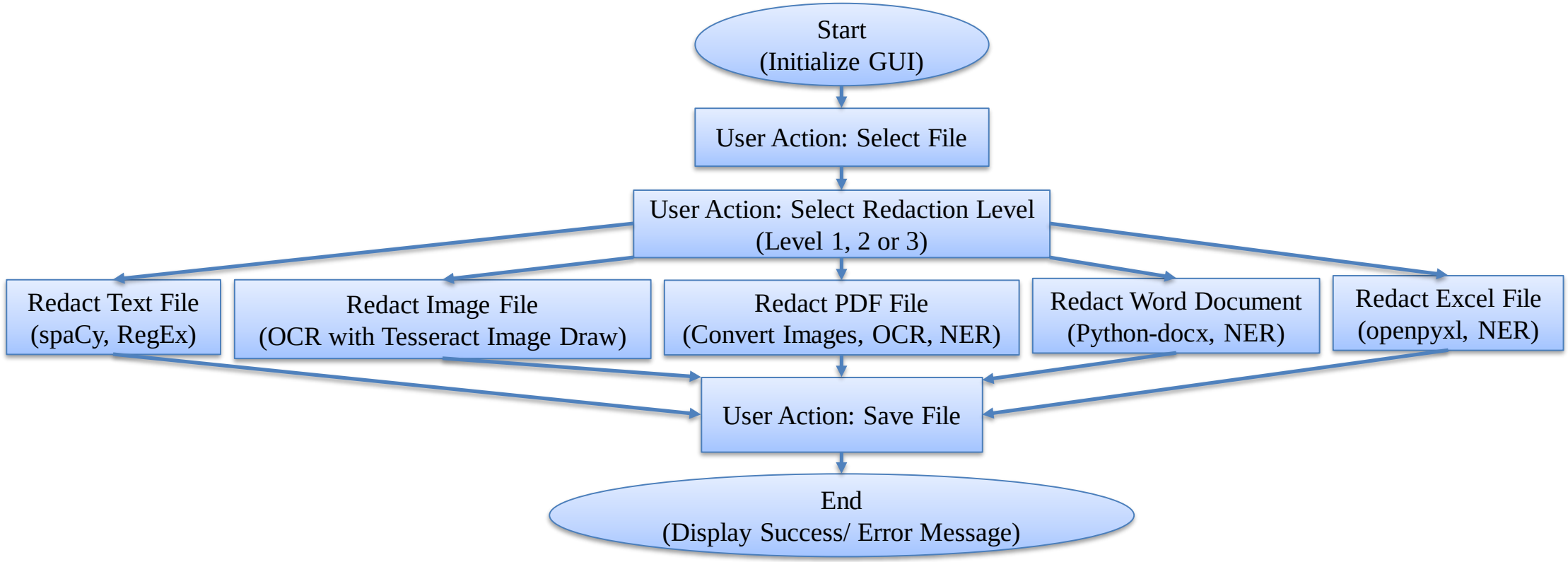
Team Rookie

TECHNICAL APPROACH

SMART INDIA HACKATHON 2024

- **Technology Used:**
  - Python, spaCy, Hugging Face Transformer, Tesseract OCR, Tkinter, OpenPyXL, Python-docx.
  - Tools and other Libraries for image processing.

- **Work Flow:**



Team Rookie

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2024

- **Feasibility Analysis:**
  - We are building our tool with widely used open-source Python library and frameworks.
  - Our tool is easily integrated with existing system.
  - All 3 major technologies used in out tool are open source, hence showing that development cost will be low.

- **Potential Challenges and Risks:**
  - Errors in the function of OCR and NER for different formats and languages.
  - Guaranteeing that all data is erased and that client's privacy is also preserved.
  - Recognition of the words of the text, images, PDF files.
  - Identification and segregation of Personal Key Identifiers from other text in file.

- **Strategies to Overcome Challenges:**
  - Uninterrupted model training as well as tuning.
  - By using strong data sanitization measures as well as employing solutions for data deletion.
  - How to use Hugging Face Transformer for Accuracy.

Team Rookie

IMPACTS AND BENEFITS

SMART INDIA HACKATHON 2024

- **Potential impact:**
  - Protects data of different organizations including government, legal, corporate, healthcare, finance sectors and etc.
  - Assists with compliance to privacy laws and regulations such as GDPR, HIPAA, and CCPA because the sensitive data is well mediated.
  - Enables organizations to encourage an open document exchange that benefits the company but is secured to safeguard against leaked information.

- **Benefits of the Solution:**
  - Multi-Format Support.
  - Customizable Redaction Levels.
  - Security Assurance.
  - Enhanced Document Integrity.
  - Adaptable to Different Use Cases.

Team Rookie

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2024

- **Research Work:**
  - Use of spaCy, Hugging Face Transformers, and Tesseract OCR.
  - Transformers for Natural Language Processing by Dennis Rothman.
  - Named Entity Relationship-NER-Paper by Pengfei Liu, Jinlan Fu.
  - Image Processing Based Extraction of Data From Graphical Representations by Viswanath Reddy.
  - Analysis report of widely used tool for redaction by Team Rookie (https://bit.ly/TeamRookie)

- **Links to References:**
  - https://github.com/pfliu-nlp/Named-Entity-Recognition-NER-Papers
  - https://github.com/tesseract-ocr/tesseract
  - https://github.com/explosion/spaCy
  - https://bit.ly/TeamRookie