**Module Title:**  
**AI-Powered Named Entity Extraction for OSINT Enrichment**

**Submitted By:**  
Muhammad Rehan Rashid

G1F22UBSCS095

BSCS-VI (C)

### ****Objective of the AI Module****

The objective of this AI module is to enhance the functionality of an OSINT (Open-Source Intelligence) Recon Dashboard by enabling intelligent extraction of meaningful information from scraped user bios. Specifically, this module identifies named entities such as people, organizations, locations, and geopolitical entities from social media profile data using Natural Language Processing (NLP).

This process plays a critical role in contextualizing scraped data, aiding in user profiling, entity linking, and risk evaluation — making the recon system more actionable and intelligence-driven.

### ****Technology Stack****

* **Programming Language:** Python
* **NLP Library:** spaCy
* **Model Used:** xx\_ent\_wiki\_sm (Multilingual NER model)
* **Custom Extension:** spaCy EntityRuler with handcrafted Pakistani names, universities, and organizations
* **Integration Method:** Executable via command-line and integrated into a full-stack Node.js-based OSINT dashboard
* **Data Files:**
  + pakistani\_names.jsonl (used for adding custom NER rules)
  + Additional JSONL files for organizations and locations

### ****Dataset and Custom Patterns****

The module uses a combination of pretrained spaCy models and handcrafted custom rules to improve accuracy for regional data. The custom data includes:

* Pakistani male and female first names
* Recognized universities (as ORG entities)
* Local cities and provinces (as GPE/LOC)
* Popular Pakistani companies (as ORG)

These are loaded using the EntityRuler pipeline component before default NER processing, enabling domain-specific improvements to entity recognition.

### ****How the AI Module Works****

1. **Input:**  
   The module accepts a text string (usually a scraped bio) via command-line argument.
2. **Entity Detection:**  
   The text is passed through the spaCy pipeline, which performs Named Entity Recognition.
3. **Custom Rule Matching:**  
   Custom patterns are matched using EntityRuler to supplement spaCy’s default model.
4. **Merging Adjacent PERSON Entities:**  
   If two consecutive entities are both labeled as PERSON, the module intelligently merges them into full names to improve accuracy.
5. **Entity Structuring:**  
   Extracted entities are grouped by label type (e.g., PERSON, ORG, GPE, LOC) and normalized (e.g., title-cased).
6. **Output:**  
   The final output is a structured JSON object containing all detected entities, which is returned to the main OSINT Dashboard for further processing.

### ****Role in Full-Stack Web Integration****

The AI module is tightly integrated into a larger OSINT Dashboard built using the MERN (MongoDB, Express.js, React.js, Node.js) stack. Here's how it fits:

* The **backend (Node.js)** scrapes user bios from platforms like Instagram, GitHub, Twitter, and Facebook.
* Once a bio is obtained, the **backend invokes this Python module** using child\_process or similar logic.
* The module processes the bio and returns a **clean, structured JSON response** of detected entities.
* These entities are then saved in the **MongoDB database** and also displayed via the **React frontend** for analyst review.

This design ensures modularity, reusability, and the ability to independently test and upgrade the AI module without affecting the frontend or database structure.

### ****Outcome and Benefits****

* Provides meaningful intelligence from noisy text data
* Improves analyst efficiency by pre-tagging bios
* Supports risk scoring and automated prioritization
* Highly relevant to cybersecurity, OSINT, and threat intelligence workflows
* Shows how AI/NLP can be integrated into real-world full-stack apps

### ****Key Takeaways****

* Demonstrated ability to combine custom rule-based and statistical NLP for domain-specific entity extraction
* Enhanced scalability and automation of data enrichment in OSINT environments
* Modular, maintainable, and production-ready AI code integrated into a live web project
* Assignment successfully meets the AI requirement in a meaningful, original, and technically sound manner