Report Tecnico: Analisi Documentale e Protezione Dati

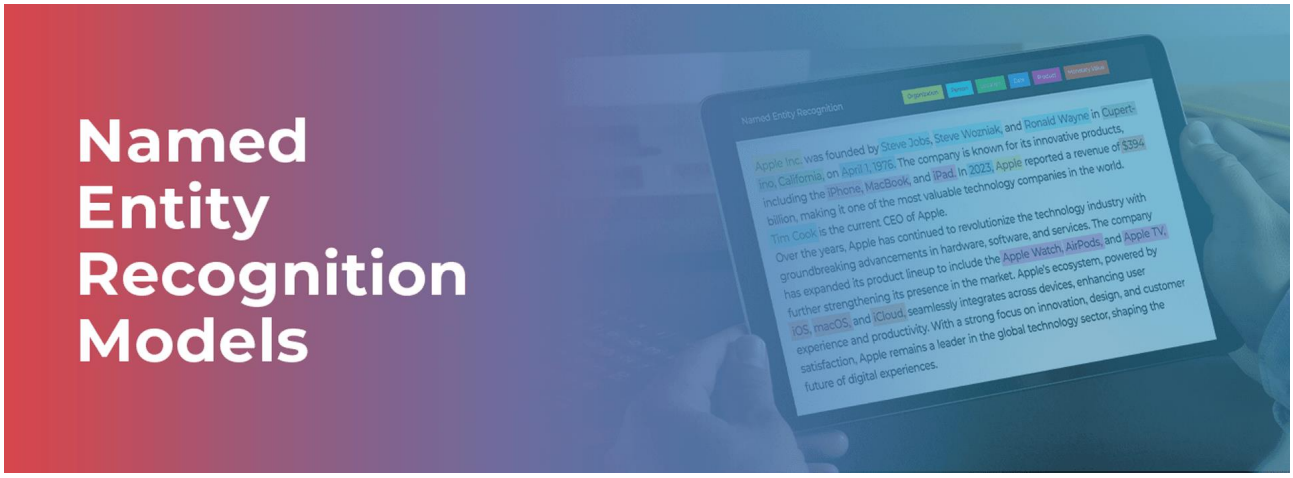

Tabella dei contenuti:

---

1. Commessa

Cliente: SmartDocs Srl
Consulente: Christian Putzu
Data: 21 giugno 2025

---

2. Premessa

SmartDocs Srl gestisce quotidianamente documenti e comunicazioni contenenti **dati sensibili** (nomi, indirizzi, IBAN, codici fiscali). Per migliorare l'efficienza operativa e garantire la conformità alle normative sulla privacy, è stato avviato un progetto volto ad **automatizzare l'estrazione di informazioni**, la **sintesi dei contenuti** e la **generazione di risposte automatiche**, preservando la **riservatezza dei dati**.

---

3. Obiettivi

- Estrarre automaticamente entità sensibili da documenti/email (NER)
- Anonimizzare i dati prima di qualsiasi elaborazione esterna
- Riepilogare i contenuti e generare risposte automatiche alle richieste dei clienti
- Minimizzare i costi di elaborazione cloud
- Garantire che i dati sensibili non escano mai dall'infrastruttura locale

---

4. Flusso di Lavoro Proposto

1. **Input:** documento/email del cliente.
2. **Estrazione e mascheramento locale:**
    - Named Entity Recognition tramite modello locale.
    - Anonimizzazione con regex e sostituzioni.
3. **Invio al cloud:**
    - Solo il testo anonimizzato viene inviato al modello GPT-4o tramite Azure AI Foundry.
4. **Output:**
    - Il modello cloud restituisce un riepilogo sintetico e una risposta automatica pronta all'uso.
    - Salvataggio in locale.

---

5. Considerazioni Tecniche Finali

Il modello locale dslim/bert-base-NER è stato testato su documenti in lingua inglese, dimostrandosi adeguato per scenari in cui è richiesta un'estrazione rapida e locale delle entità, mantenendo i dati sensibili all'interno dell'infrastruttura aziendale.

Infatti, la sua dimensione contenuta lo rendono particolarmente indicato per flussi che richiedono efficienza e controllo diretto sul trattamento dei dati per la privacy.

Tuttavia, per ottenere prestazioni superiori in termini di accuratezza e copertura delle entità, è fortemente consigliato valutare l'adozione di modelli più potenti da ospitare in locale qualora si disponga di un'infrastruttura con adeguata potenza computazionale, oppure effettuare un fine-tuning del modello stesso su un dataset specifico per il dominio d'interesse.

In alternativa, se l'infrastruttura tecnologica e la struttura economico-finanziaria aziendali lo consentono, è possibile delegare l'estrazione delle entità a soluzioni NER cloud come Azure Language, che offrono capacità più avanzate built-in facilmente implementabili e la possibilità di effettuare labeling e training personalizzati, nonché valutazione delle performance di test e deploy.
