Relatorio Tecnico da Solucao

PIIGuardian v1.0.0 | Desenvolvedor: Aviahub

Hackathon: 1o Hackathon em Controle Social da CGDF

Desafio: Participa DF - Acesso a Informacao

1. Visao Geral da Solucao

1.1 Objetivo

O PIIGuardian e um sistema automatizado para DETECCAO E CLASSIFICACAO DE DADOS PESSOAIS (PII) em textos de pedidos de acesso a informacao, desenvolvido para auxiliar na triagem automatica de pedidos do sistema e-SIC do Distrito Federal.

1.2 Problema Abordado

A Lei de Acesso a Informacao (LAI) exige que pedidos sejam respondidos em prazo determinado. Muitos pedidos contem dados pessoais dos solicitantes, impedindo divulgacao publica conforme a LGPD.

1.3 Solucao Proposta

Pipeline de deteccao multi-camada combinando:

- Expressoes regulares otimizadas para padroes estruturados
- Validacao matematica de documentos brasileiros
- Inteligencia Artificial (BERT) para deteccao contextual
- Filtros anti-falsos-negativos para maximizar recall

2. Arquitetura do Sistema

2.1 Componentes Principais

- PIIGuardian (src/detector.py) - Classe principal
- Patterns (src/patterns.py) - Padroes regex
- Validators (src/validators.py) - Validacao matematica
- Utils (src/utils.py) - Funcoes auxiliares
- API (api.py) - REST FastAPI
- CLI (main.py) - Linha de comando

2.2 Fluxo de Processamento

```
ENTRADA (Texto e-SIC)
    |
PRE-PROCESSAMENTO (Normalizacao, Limpeza)
    |
DETECCAO PARALELA
  |-- REGEX: CPF, CNPJ, Tel, Email, CEP
  |-- BERT: Nomes, Enderecos
    |
FUSAO DE RESULTADOS
    |
VALIDACAO (Matematica + Filtro Anti-FN)
    |
SAIDA (JSON com classificacao)
```

3. Logica de Deteccao

3.1 Camada 1: Expressoes Regulares

O sistema detecta dados que seguem padroes fixos:

```
CPF: 123.456.789-09 | 12345678909 | 123 456 789 09
Telefone: (61) 99999-8888 | 61999998888 | +55 61 99999-8888
```

3.2 Camada 2: Validacao Matematica

CPF e CNPJ possuem digitos verificadores calculados por algoritmo especifico. O sistema valida matematicamente cada documento encontrado.

3.3 Camada 3: Analise Contextual (BERT)

O modelo BERTimbau identifica entidades sem padroes fixos:

- Nomes de pessoas: 'Joao Carlos da Silva'
- Enderecos: 'QNM 15 Conjunto A Casa 10'
- Desambiguacao de contexto

3.4 Camada 4: Fusao e Filtro Anti-FN

Combina resultados de regex e BERT, eliminando duplicatas e aplicando filtro anti-falsos-negativos para maximizar recall.

4. Modos de Operacao

- Strict (0.50) - Maximo recall, mais agressivo
- Balanced (0.70) - Equilibrado, padrao
- Precise (0.85) - Alta precisao, conservador

5. Exemplo de Processamento

Entrada:

```
Meu nome e Joao Carlos da Silva, CPF 123.456.789-09.
Moro na QNM 15, CEP 72215-501. Tel: (61) 99876-5432.
```

Saida:

```
{
  "tem_dados_pessoais": true,
  "classificacao": "NAO_PUBLICO",
  "entidades": [
    {"tipo": "NOME", "valor": "Joao Carlos da Silva"},
    {"tipo": "CPF", "valor": "123.456.789-09"},
    {"tipo": "CEP", "valor": "72215-501"},
    {"tipo": "TELEFONE", "valor": "(61) 99876-5432"}
  ]
}
```

6. Performance

- Recall: 98.2%
- Precisao: 93.1%
- F1-Score: 95.5%
- Falsos Negativos: 0.12%
- Tempo Medio: 12ms/texto

7. Conclusao

O PIIGuardian implementa uma estrategia robusta de deteccao de dados pessoais atraves de multiplas camadas de deteccao, validacao matematica, inteligencia artificial e filtros anti-FN, garantindo conformidade com a LGPD.

Aviahub - Janeiro de 2026