

Relatório Técnico: Desenvolvimento de Chatbot Analítico com Agentes de IA e Dados Governamentais

Luiz Carlos Porto do Carmo

1. Introdução: Este projeto apresenta o desenvolvimento de um *chatbot* inteligente projetado para responder a perguntas em linguagem natural, fundamentadas em um conjunto de dados real do governo brasileiro. O *dataset* alvo, denominado PORTES_2025, compila registros de portes de arma de fogo na categoria “Defesa Pessoal”, abrangendo variáveis como unidade federativa (UF), município, espécie da arma, calibre, sexo do requerente, marca e status do porte.

A proposta integra conceitos avançados de Mineração de Dados (*Data Mining*) e Engenharia de Dados com a utilização de Agentes de IA. A solução foi implementada utilizando a biblioteca Agno para orquestração de agentes e o modelo Gemini 2.0 Flash (Google) para processamento de linguagem natural. O sistema é capaz de interpretar a intenção do usuário, converter linguagem natural em consultas estruturadas e extrair informações precisas do conjunto de dados.

2. Arquitetura do Sistema e Decisões Técnicas: A arquitetura foi concebida para ser modular, escalável e de fácil manutenção, dividida em três pilares fundamentais:

- 2.1. Ingestão e Tratamento de Dados (ETL) O carregamento dos dados foi realizado via biblioteca Pandas, com uma etapa robusta de pré-processamento para mitigar inconsistências típicas de bases governamentais:
 - Codificação: Tratamento de *encoding* (latin1) para suporte a caracteres especiais.
 - Delimitadores: Ajuste de leitura para arquivos CSV separados por ponto e vírgula (;).
 - Normalização: Padronização de nomes de colunas e tipagem de dados (conversão da coluna TOTAL para numérico).
- 2.2. Implementação da Ferramenta de Consulta (Tooling) Desenvolveu-se a função `consultar_portes`, decorada como uma *Tool* no ecossistema Agno. Isso permite que o Agente invoque a função programaticamente. A ferramenta suporta filtros dinâmicos e simultâneos (UF, Município, Sexo, Espécie, Marca, Calibre, Ano) e realiza operações de agregação e *ranking* diretamente no *DataFrame*, transformando a intenção do usuário em operações vetoriais de dados.
- 2.3. Agente de IA (Agno + Gemini) O núcleo da inteligência utiliza o modelo Gemini 2.0 Flash, configurado através do framework Agno. O agente atua como uma ponte entre o Processamento de Linguagem Natural (NLP) e a análise quantitativa, sendo responsável por:
 - Identificar a necessidade de consulta externa.
 - Extrair entidades da pergunta (ex: "Bahia" -> UF= ' BA ').
 - Manter o contexto conversacional.
 - Sintetizar os resultados brutos da *Tool* em respostas humanizadas.

3. Fluxo de Funcionamento A interação ocorre via terminal, seguindo um fluxo lógico de processamento:

1. Input: O usuário insere uma pergunta (ex: "Qual o calibre mais usado?").
2. Inferência: O Agente analisa a semântica e decide invocar a ferramenta `consultar_portes`.
3. Execução: A *Tool* aplica os filtros no *dataset* carregado em memória e retorna os dados brutos ou sumarizados.
4. Resposta: O Agente interpreta o retorno da função e gera uma resposta final em português fluido.

O sistema suporta inferências como contagem total, *rankings* (top N), filtragens combinadas e comparações.

4. Resultados e Exemplos de Uso O sistema demonstrou alta precisão na interpretação de consultas variadas:

- Consulta Quantitativa:
 - *Pergunta*: "Quantos portes foram emitidos na Bahia?"
 - *Resposta*: "Foram encontrados 1.284 portes de arma com os filtros informados (UF = BA)."
- Consulta de Ranking:
 - *Pergunta*: "Qual o calibre mais comum no dataset?"
 - *Resposta*: "Ranking por CALIBRE_ARMA (top 5): .380 ACP (7.142), .38 TPC (1.832)..."
- Consulta Segmentada:
 - *Pergunta*: "Quantos portes foram emitidos para mulheres?"
 - *Resposta*: "Foram encontrados 312 portes emitidos para o sexo feminino."

5. Limitações e Trabalhos Futuros Embora funcional, o protótipo possui limitações identificadas:

- Sensibilidade do ETL: Alterações estruturais no CSV (mudança de delimitadores ou nomes de colunas) exigem refatoração do código.
- Ambiguidade Semântica: Consultas extremamente vagas podem gerar interpretações equivocadas pelo modelo.
- Persistência: O contexto é mantido apenas em memória (sessão), sem histórico persistente entre execuções.
- Dependência de API: A operação depende integralmente da disponibilidade e autenticação da API do Google Gemini.

6. Conclusão O projeto validou a eficácia da combinação entre IA Generativa e ferramentas de análise de dados estruturados. O *chatbot* transcendeu a simples geração de texto, atuando como um analista de dados autônomo capaz de extrair *insights* do arquivo PORTES_2025. O desenvolvimento consolidou conhecimentos em orquestração de agentes (Agno), manipulação de dados em larga escala (Pandas) e integração com LLMs de última geração.