ROTEIRO DE APRESENTAÇÃO ORAL (5-10 minutos)

---

1. INTRODUÇÃO (30 segundos)

"Meu trabalho comparou dois modelos de Question Answering: o DistilBERT e o RoBERTa. Usei 1.000 perguntas do dataset DBpedia e analisei qual modelo seria melhor para usar em produção."

---

2. MÉTRICAS QUANTITATIVAS (1 minuto)

"Primeiro, calculei as métricas básicas:

- As perguntas têm em média 4.80 palavras
- O DistilBERT teve score médio de 0.41, o RoBERTa 0.33
- O overlap médio foi 83% pro DistilBERT e 78% pro RoBERTa

Isso já mostra que o DistilBERT tem mais confiança e extrai mais do texto original."

---

3. ANÁLISE DO SCORE vs QUALIDADE (1 minuto)

"Mas aí eu fui verificar: será que score alto significa resposta boa?

Descobri que:

- Score muito alto (tipo 0.99) geralmente é resposta certa sim
- Score muito baixo (tipo 0.003) significa que o modelo sabe que não sabe
- Mas no meio do caminho, não dá pra confiar cegamente

Por exemplo: uma pergunta teve score 0.98, mas a resposta estava errada. Então score alto não garante nada sozinho."

---

4. ANÁLISE DO OVERLAP (1 minuto)

"O overlap mede quanto da resposta está no texto original.

- Overlap alto (acima de 90%): resposta veio direto do texto, mais seguro
- Overlap baixo (abaixo de 50%): pode ser alucinação

O DistilBERT teve só 7.7% de casos com overlap baixo. O RoBERTa teve 12.2%. Ou seja, o RoBERTa 'inventa' mais."

---

5. ANÁLISE QUALITATIVA - TOP 10 (1 minuto)

"Analisei manualmente os 10 exemplos com maior score.

Eram perguntas tipo 'quem fundou tal empresa', 'qual a nacionalidade de fulano', 'em que ano aconteceu isso'.

Todas as respostas do DistilBERT estavam certas. Perguntas factuais simples ele manda muito bem."

---

6. ANÁLISE QUALITATIVA - BOTTOM 10 (1 minuto)

"Nos 10 com menor score, eram perguntas mais difíceis:

- 'Quais livros Tolkien escreveu' - tem vários, não só um
- 'População do Quirguistão' - o texto não tinha esse dado

Aí percebi que o modelo sabe quando não sabe. O score baixo é um aviso."

---

7. ANÁLISE QUALITATIVA - DIVERGENTES (2 minutos)

"Nos 5 casos onde os modelos discordaram, analisei quem estava mais certo.

O caso mais grave foi a pergunta 'youngest female chess player'. O RoBERTa respondeu 'Bobby Fischer', que é HOMEM! Isso é uma alucinação séria.

No geral, empatou 2 a 2, mas os erros do RoBERTa foram mais graves."

---

8. CONCLUSÃO (1 minuto)

"Por isso escolhi o DistilBERT para produção.

Não foi só por causa dos números. Foi porque:

1. No Top 10, ele acertou tudo
2. No Bottom 10, ele sinalizou incerteza corretamente
3. Nos Divergentes, quando o RoBERTa errou, errou feio

Claro que em produção real precisaria de filtros e revisão humana, mas entre os dois, o DistilBERT é mais confiável."

PERGUNTAS QUE A PROFESSORA PODE FAZER:

P: Por que você escolheu esses dois modelos? R: "O DistilBERT é uma versão compacta do BERT, mais

rápido e leve. O RoBERTa é uma versão otimizada, teoricamente mais robusta. Quis comparar um modelo leve com um mais pesado."

P: O que é overlap? R: "É a porcentagem de palavras da resposta que aparecem no texto original. Se o overlap é 100%, toda a resposta foi copiada do texto. Se é 0%, o modelo inventou tudo."

P: O que é alucinação em QA? R: "É quando o modelo gera uma resposta que não está no texto e não faz sentido. Tipo responder 'Bobby Fischer' pra uma pergunta sobre a jogadora de xadrez mais jovem - ele é homem."

P: Por que score alto não garante resposta certa? R: "Porque o score é a confiança do modelo, não a certeza. O modelo pode estar muito confiante numa resposta errada. Por isso precisamos de outras métricas como overlap."

P: Você usaria esse modelo em produção de verdade? R: "Sim, mas com filtros. Só aceitaria respostas com score acima de 0.30 e overlap acima de 0.70. E casos intermediários iriam pra revisão humana."

TERMOS TÉCNICOS PRA LEMBRAR:

- **Score**: Confiança do modelo na resposta (0 a 1)
- **Overlap**: Sobreposição entre resposta e texto original
- **Alucinação**: Quando o modelo inventa informação
- **Ground Truth**: Resposta correta de referência (não tínhamos no dataset)
- **Extração**: O modelo pega um trecho do texto como resposta
- **Inferência**: O modelo deduz algo que não está explícito