RESUMO PARA APRESENTAÇÃO DO TRABALHO

Avaliação de Modelos de Question Answering

🎯 O QUE EU FIZ NESTE TRABALHO

Comparei dois modelos de Question Answering do Hugging Face usando 1.000 perguntas do dataset DBpedia:

- **Modelo 1**: DistilBERT (distilbert-base-cased-distilled-squad)
- **Modelo 2**: RoBERTa (deepset/roberta-base-squad2)

📊 CRITÉRIO A - TAMANHO MÉDIO DAS PERGUNTAS

O que fiz: Calculei quantas palavras tem cada pergunta em média.

Resultado: As perguntas têm em média **4.80 palavras**.

Por que isso importa: Perguntas curtas como "who is X" ou "what year" tendem a ter respostas mais diretas. Perguntas maiores costumam ser mais complexas.

📊 CRITÉRIO B - SCORE MÉDIO E QUALIDADE

O que fiz: Calculei a confiança média de cada modelo e analisei se essa confiança reflete a qualidade real.

Resultados:

- DistilBERT: score médio de **0.4131**
- RoBERTa: score médio de **0.3310**

Minha análise sobre score vs qualidade:

O score é parcialmente confiável:

- **Scores muito altos (>0.95)** → geralmente a resposta está correta e explícita no texto
- **Scores muito baixos (<0.02)** → o modelo está incerto, e faz sentido porque a pergunta é ambígua ou não tem resposta clara
- **Scores intermediários** → não dá pra confiar cegamente, precisa verificar

Exemplo que comprova isso: A pergunta "who played tarzan in the movies" teve score 0.98, mas a resposta foi "Tarzan" (o personagem, não o ator). Ou seja, score alto não garante resposta certa.

📊 CRITÉRIO C - OVERLAP (SOBREPOSIÇÃO)

O que fiz: Medi quanto das palavras da resposta aparecem no texto original.

Resultados:

- DistilBERT: overlap médio de **83.71%**
- RoBERTa: overlap médio de **78.41%**

O que isso significa:

Overlap	Interpretação
Alto (>90%)	Resposta copiada direto do texto - baixo risco de erro
Médio (50-90%)	Resposta parcial - precisa verificar
Baixo (<50%)	Possível alucinação - o modelo pode ter inventado

Comparação dos modelos:

- DistilBERT teve **7.7%** de casos com overlap baixo (possível alucinação)
- RoBERTa teve **12.2%** de casos com overlap baixo

Conclusão: DistilBERT é mais conservador e erra menos "inventando" coisas.

📊 CRITÉRIO D - ANÁLISE QUALITATIVA DOS 25 EXEMPLOS

O que fiz: Analisei manualmente 25 exemplos específicos:

- 10 com maior score (onde o modelo tinha mais confiança)
- 10 com menor score (onde o modelo tinha menos confiança)
- 5 onde os modelos deram respostas diferentes

MEUS ACHADOS NO TOP 10 (MAIOR SCORE):

Analisando os exemplos com score mais alto, percebi que:

- Perguntas do tipo "who is", "what year", "what nationality" funcionam muito bem
- Quando a resposta é um nome, data ou fato único, o modelo acerta
- Exemplo: "who started labatt beer" → "John Kinder Labatt" (correto!)
- Exemplo: "what year was the jacquard loom invented" → "1801" (correto!)

Minha conclusão: O DistilBERT é excelente para perguntas factuais simples.

MEUS ACHADOS NO BOTTOM 10 (MENOR SCORE):

Analisando os exemplos com score mais baixo, percebi que:

- São perguntas abertas ou sem resposta única
- Exemplo: "what books did tolkien write" - tem vários livros, não só um
- Exemplo: "kyrgyz population" - o texto não tinha o número exato

Minha conclusão: O modelo sabe quando não sabe! O score baixo é um aviso de "não confie muito em mim aqui".

MEUS ACHADOS NOS 5 DIVERGENTES:

Quando os modelos discordaram, analisei quem estava mais certo:

1. **"what is the jazz guitar"**
    - DistilBERT: deu a definição (correto)
    - RoBERTa: falou de amplificação (detalhe técnico, não a definição)
    - **Vencedor**: DistilBERT
2. **"what river is javari located in"**
    - DistilBERT: "The Javary River" (é o próprio rio)
    - RoBERTa: "the Amazon" (é a bacia, não onde está)
    - **Vencedor**: DistilBERT
3. **"youngest female chess player"**
    - DistilBERT: resposta incompleta
    - RoBERTa: "Bobby Fischer" (que é HOMEM!)
    - **Vencedor**: Nenhum, mas RoBERTa errou feio
4. **"who makes jupiter computer"**
    - DistilBERT: resposta confusa
    - RoBERTa: identificou que é uma empresa
    - **Vencedor**: RoBERTa
5. **"what are jewish holidays"**
    - DistilBERT: resposta redundante
    - RoBERTa: mais direto
    - **Vencedor**: RoBERTa

Placar final dos divergentes: 2x2, com 1 empate. Mas os erros do RoBERTa foram mais graves (como errar o gênero da pessoa).

🏆 MINHA ESCOLHA PARA PRODUÇÃO: DISTILBERT

Por que escolhi o DistilBERT?

Baseado na minha análise dos 25 exemplos (não só nas métricas):

1. **No Top 10**: Acertou todas as perguntas factuais que analisei
2. **No Bottom 10**: Sinalizou corretamente quando não tinha certeza
3. **Nos Divergentes**: Quando o RoBERTa errou, errou feio (ex: Bobby Fischer pra pergunta sobre mulher)
4. **Menos alucinações**: 7.7% vs 12.2% de casos problemáticos

Quando eu NÃO usaria o DistilBERT:

- Perguntas abertas tipo "o que é..." ou "defina..."
- Quando preciso de respostas mais interpretativas
- Nesses casos, talvez um modelo maior ou generativo fosse melhor

---

💡 O QUE EU APRENDI COM ESTE TRABALHO

1. **Score alto não garante resposta certa -** vi exemplos de score 0.98 com resposta errada
2. **Overlap é um bom indicador de confiabilidade -** se a resposta não está no texto, desconfie
3. **Modelos menores podem ser melhores -** o DistilBERT é menor e mais rápido, mas foi mais confiável que o RoBERTa no meu dataset
4. **Análise qualitativa é essencial -** só olhando os números eu não teria percebido o erro grave do Bobby Fischer
5. **Em produção, precisa de filtros -** não dá pra confiar cegamente no modelo, precisa de thresholds e revisão humana

---

📈 RESUMO DOS NÚMEROS

Métrica	DistilBERT	RoBERTa
Score Médio	0.4131	0.3310
Overlap Médio	83.71%	78.41%
Casos com overlap ≥90%	62.6%	53.3%
Casos com overlap <50% (risco)	7.7%	12.2%

Métrica	DistilBERT	RoBERTa
Tamanho médio das perguntas	4.80 palavras	4.80 palavras

✅ CHECKLIST DO QUE ENTREGUEI

- [x] 1.000 exemplos do shard_007
- [x] Dois modelos diferentes do Hugging Face
- [x] Tamanho médio das perguntas calculado
- [x] Score médio calculado e analisado
- [x] Overlap calculado e interpretado
- [x] 10 exemplos Top analisados manualmente
- [x] 10 exemplos Bottom analisados manualmente
- [x] 5 exemplos Divergentes analisados manualmente
- [x] Análise de contexto, correção e alucinação
- [x] Escolha de modelo para produção fundamentada na análise qualitativa