Metodologia de Métricas de Avaliação

Visao Geral

Para avaliar se um modelo esta respondendo corretamente, comparamos suas respostas com um **GT de transcrições disponibilizado.**

Um modelo pode acertar o diagnostico usando palavras diferentes do gabarito. Por exemplo, o GT pode dizer "AVC isquemico" e o modelo responder "Acidente Vascular Cerebral do tipo isquemico" - ambos estao corretos, mas as palavras sao diferentes.

Para resolver isso, usamos **similaridade semantica**: Comparando o *significado* dos textos, não apenas as palavras exatas.

Exemplo pratico:

- "Fratura no femur" -> [0.82, 0.15, 0.43, ...]
- "Osso da coxa quebrado" -> [0.79, 0.18, 0.41, ...] (muito similar!)
- "Exame de sangue normal" -> [0.12, 0.67, 0.23, ...] (bem diferente)

Para gerar esses vetores, usamos o modelo *all-mpnet-base-v2* do Sentence Transformers, que foi treinado especificamente para entender similaridade entre textos. Depois, calculamos o quanto dois vetores "apontam na mesma direção" (similaridade de cosseno).

As 4 Métricas que Usamos

Avaliamos cada resposta em 4 dimensoes diferentes, cada uma capturando um aspecto importante:

1. Similaridade Geral (peso: 20%)

O que mede: O alinhamento geral entre a resposta do modelo e o gabarito completo.

Como funciona: Pegamos todo o conteudo do GT (diagnostico + descricao) e comparamos com toda a resposta do modelo. E como perguntar: "no geral, o modelo falou sobre a mesma coisa que o gabarito?"

Exemplo: Se o GT fala sobre "glaucoma com aumento de pressao intraocular" e o modelo responde sobre "pressao elevada no olho causando dano ao nervo optico", a similaridade sera alta (~80%), mesmo sem usar as mesmas palavras.

2. Diagnostico Semantico (peso: 25%)

O que mede: Se o modelo identificou corretamente a condicao/patologia, mesmo usando termos diferentes.

Como funciona: Pegamos o diagnostico do GT e procuramos na resposta do modelo a frase que mais se aproxima semanticamente. Dividimos a resposta em frases e encontramos a "melhor correspondencia".

Por que e importante: O diagnostico e a informacao mais critica. Um modelo pode descrever bem a imagem mas errar o diagnostico, ou vice-versa. Esta metrica foca especificamente no acerto do diagnostico.

Exemplo: GT diz "Retinopatia diabetica proliferativa". O modelo responde "A imagem de fundo de olho mostra sinais compativeis com complicacao ocular do diabetes, com neovascularizacao". Mesmo sem dizer "retinopatia diabetica", a similaridade semantica sera alta.

3. Tipo de Imagem (peso: 15%)

O que mede: Se o modelo identificou corretamente o tipo de exame (OCT, raio-X, tomografia, etc).

Como funciona: Verificamos se os termos tecnicos de tipo de exame presentes no GT tambem aparecem na resposta do modelo. Buscamos por termos como: OCT, tomografia, raio-X, ressonancia, ultrassom, endoscopia, etc.

Por que e importante: Identificar corretamente o tipo de exame e fundamental. Um erro aqui pode indicar que o modelo nao entendeu o que esta analisando. Por exemplo, confundir uma tomografia com um raio-X pode levar a interpretacoes completamente erradas.

4. Cobertura de Fatos (peso: 15%)

O que mede: Quantas informacoes clinicas relevantes do gabarito o modelo incluiu na resposta.

Como funciona: Comparamos a descricao detalhada do GT com os achados reportados pelo modelo. Quanto mais informacoes relevantes o modelo mencionar, maior o score.

Exemplo: Se o GT menciona "nodulo de 2cm no lobo direito, margens irregulares, realce heterogeneo", e o modelo responde "lesao nodular no lobo direito com bordas irregulares", ele cobriu 2 de 3 fatos principais (nodulo + margens), resultando em boa cobertura.

Score Final: Combinando Tudo

O score final e uma **média ponderada** das 4 métricas:

Metrica	Peso
Similaridade Geral	20%
Diagnostico Semantico	25%
Tipo de Imagem	15%
Cobertura de Fatos	15%

Nota: Os pesos priorizam o diagnostico (25%) pois é a informação mais crítica. Tipo de imagem e cobertura de fatos tem pesos menores (15% cada) mas ainda são importantes para garantir uma avaliacão completa.

Interpretando os Resultados

Os valores de similaridade semantica foram ajustados para uma escala mais intuitiva:

Score	Interpretacao
< 50%	Baixo - resposta pouco alinhada com o GT
50% - 65%	Moderado - captura parte do conteudo esperado
65% - 80%	Bom - resposta bem alinhada com o GT
> 80%	Excelente - alta fidelidade ao gabarito

Resumo

Nossa metodologia avalia os modelos de IA em 4 dimensoes complementares, usando similaridade semantica para capturar acertos mesmo quando as palavras sao diferentes. O score final combina essas dimensoes com pesos que priorizam o diagnostico correto.

Essa abordagem permite comparar modelos de forma objetiva e identificar qual oferece o melhor equilibrio entre qualidade de resposta e tempo de processamento.