Relatório de Análise de Similaridade entre Modelos

Seu Nome

March 29, 2025

Contents

Introdução

Este relatório apresenta a análise comparativa dos modelos baseados nas métricas de similaridade entre respostas. Os dados foram processados a partir de um arquivo pickle que contém as respostas dos modelos e as respostas de referência (ground truth). Diversas métricas de similaridade (como Cosine Similarity, Difflib, BERTScore, entre outras) foram calculadas para cada comparação. O objetivo é entender a performance dos modelos a partir de diferentes perspectivas estatísticas e correlacionais.

Metodologia

Inicialmente, os dados foram carregados e convertidos para um formato de dicionário. Foram construídos dois DataFrames principais: **df_overall**: contém as estatísticas gerais (overall similarity) para cada questão e modelo; **df_detailed**: armazena, para cada comparação (modelo versus ground truth), os valores de diversas métricas de similaridade. Em seguida, foram aplicadas análises descritivas, de correlação e variação, além da exportação dos dados para arquivos CSV e tabelas em LaTeX. Os gráficos gerados (barras, histogramas, boxplots, heatmaps e scatter plots) ilustram visualmente os resultados e auxiliam na interpretação dos dados.

Resultados

Análise Descritiva da Similaridade Geral

A tabela a seguir (Tabela 1) apresenta as estatísticas descritivas da overall similarity por modelo, incluindo média, mediana, desvio padrão, valor mínimo e máximo.

Table 1: Estatísticas de Overall Similarity por Modelo

$model_name$	mean	median	std	min	max
gemini-1.5-flash	0.624493	0.667514	0.243642	0.000000	1.000000
gemini-1.5-flash-8b	0.695548	0.734200	0.284632	0.000000	1.000000
gemini-1.5-pro	0.602949	0.657294	0.196461	0.000000	0.806889
gemini-2.0-flash	0.704191	0.741846	0.254234	0.000000	1.000000
gemini-2.0-flash-exp	0.712266	0.735477	0.250689	0.000000	1.000000
gemini-2.0-flash-lite	0.661387	0.659615	0.257016	0.000000	1.000000
gemini-2.0-flash-thinking-exp-1219	0.719972	0.741123	0.247354	0.000000	1.000000

Frequência de Casos com Scores Vazios

A tabela a seguir (Tabela 2) mostra a contagem e o percentual de casos com dicionários de scores vazios para cada modelo.

Table 2: Frequência de Casos com Scores Vazios por Modelo

$model_name$	empty_counts	empty_percent
gemini-1.5-flash	3	7.320000
gemini-1.5-flash-8b	5	12.200000
gemini-1.5-pro	2	4.880000
gemini-2.0-flash	3	7.320000
gemini-2.0-flash-exp	3	7.320000
gemini-2.0-flash-lite	3	7.320000
gemini-2.0-flash-thinking-exp-1219	2	4.880000

Estatísticas Agregadas das Métricas Detalhadas

A tabela a seguir (Tabela 3) apresenta as estatísticas agregadas das métricas detalhadas (média, mediana, desvio padrão, mínimo e máximo) para cada modelo.

Table 3: Estatísticas Agregadas das Métricas Detalhadas por Modelo

	Difflib Similarity	Difflib Similarity.1	Difflib Similarity.2	Difflib Similarity.
NaN	mean	median	std	min
model_name	NaN	NaN	NaN	NaN
gemini-1.5-flash	0.2799203811434493	0.21922847564710055	0.24785096376552662	0.0
gemini-1.5-flash-8b	0.42941601327943457	0.2646458758379481	0.36001778190881295	0.0
gemini-1.5-pro	0.20687483195529055	0.2087912087912088	0.14091516134442492	0.00607902735562
gemini-2.0-flash	0.44854865960863294	0.39603729603729604	0.3249836145276488	0.0
gemini-2.0-flash-exp	0.47128847317169426	0.4190522557611165	0.32206984292687246	0.0
gemini-2.0-flash-lite	0.39852259018231834	0.30897877223178427	0.3110210418287233	0.0
gemini-2.0-flash-thinking-exp-1219	0.44947587816004714	0.3649122807017544	0.32906983271331397	0.0

Discussão

A análise dos dados indica que o modelo **gemini-2.0-flash-thinking-exp-1219** apresentou a maior média de overall similarity (0.72), enquanto o modelo **gemini-1.5-pro** apresentou a menor média (0.60). Essa diferença ressalta variações na performance dos modelos em aderência às respostas de referência.

Adicionalmente, observa-se que o percentual de casos com dicionário de scores vazio foi mais elevado para o modelo **gemini-1.5-flash-8b** (12.2

A análise das estatísticas agregadas das métricas detalhadas evidencia variações na consistência dos modelos, o que pode ser explorado para identificar possíveis ajustes nos algoritmos de resposta.

Conclusões

Com base nos dados analisados, pode-se concluir que:

- Modelos com maiores médias de overall similarity tendem a aderir melhor às respostas de referência, embora a variabilidade dos scores deva ser considerada.
- O elevado percentual de casos com scores vazios em alguns modelos indica a necessidade de investigar a qualidade dos dados e o processo de geração de scores.
- As estatísticas agregadas das métricas detalhadas oferecem insights sobre a consistência dos modelos, o que pode orientar ajustes para aprimorar a performance.

Anexos

Os anexos deste relatório incluem as tabelas geradas a partir dos arquivos CSV: `overall_stats.csv`, `empty_scores_frequency.c` e `aggregated_metrics.csv`, além de todas as imagens geradas (gráficos em formato PNG presentes na pasta `analysis_results`).

Referências

Bibliotecas utilizadas:

- **pandas**, **numpy**: para manipulação e análise dos dados.
- **matplotlib**, **seaborn**: para geração dos gráficos.
- **PyLaTeX**: para a criação deste relatório em LaTeX.

Além disso, é necessário ter instalada uma distribuição LaTeX (como TeX Live ou MiKTeX) para compilar o documento.