UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIX SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2025

Uso de LLMs para apoiar análise qualitativa de dados na investigação de fatores humanos em engenharia de software

Gabriel Cordeiro Moraes[1]; José Amâncio Macedo Santos[2]

1. Bolsista – Modalidade Bolsa/PVIC, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: gcmorais@ecomp.uefs.br

2. Orientador, Departamento de Tecnologia, Universidade Estadual de Feira de Santana, e-mail: zeamancio@uefs.br

PALAVRAS-CHAVE: LLM; Engenharia de Software; Análise Qualitativa.

INTRODUÇÃO

A análise qualitativa de dados é uma atividade que demanda esforço significativo por parte dos pesquisadores (Bijker et al., 2024). Com o avanço dos modelos de linguagem de grande escala (Large Language Models – LLMs), as Inteligências Artificiais (IAs) têm sido cada vez mais exploradas no apoio em tarefas de processamento de linguagem natural, incluindo a análise qualitativa (Xiao et al., 2023). Um campo promissor para essa aplicação é o estudo de fatores humanos na Engenharia de Software, área que frequentemente requer análises qualitativas aprofundadas (Capretz, 2014).

Recentemente, alguns estudos disponibilizaram conjuntos de dados qualitativos contendo códigos e codebooks relacionados a temas como empatia e segurança psicológica na engenharia de software (Cerqueira et al., 2024; Santana et al., 2024). Esses dados foram produzidos a partir de métodos tradicionais de análise qualitativa, conduzidos exclusivamente por pesquisadores humanos. Assim, esses conjuntos podem ser usados como referência para avaliar em que medida os modelos de linguagem de grande escala (LLMs) são capazes de replicar a classificação e codificação realizadas manualmente por especialistas humanos, contribuindo para o desenvolvimento de ferramentas para apoiar o avanço da compreensão sobre fatores humanos na Engenharia de Software.

Por estes motivos, o presente trabalho visa avaliar a capacidade de modelos de linguagem de grande escala (LLMs) de replicar a classificação e a codificação de dados qualitativos realizadas por analistas humanos em estudos sobre fatores humanos na Engenharia de Software. Além de investigar as contribuições das IAs para as pesquisas relacionadas à análise de sentimentos e fatores humanos.

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

O estudo realizado neste trabalho utiliza o conjunto de dados disponibilizado por Cerqueira et al. (2023), contendo códigos e codebooks relacionados aos temas de

empatia na Engenharia de Software. Em seguida, foi realizada a preparação dos dados, envolvendo a limpeza e organização dos trechos de textos disponibilizados no referido conjunto de dados para garantir compatibilidade com a entrada dos modelos LLMs. Foi utilizado o Gemini na versão pro 2.5 como um modelos LLM, e técnicas de engenharia de prompt zero-shot e few-shot para as análises.

As técnicas de engenharia de prompt incluíram prompts diretos (zero-shot) e prompts com exemplos contextuais (few-shot), que foram refinados através de diferentes execuções do estudo. Para cada técnica selecionada, foram criados e revisados prompts buscando maximizar a clareza e a adequação das respostas geradas pelo modelo. Após o desenvolvimento dos prompts, foi realizada a execução do experimento, na qual os textos qualitativos foram submetidos ao modelo utilizando cada técnica de prompt definida. Para execução do experimento foi desenvolvido um script em linguagem Python (um agente) para interagir com a LLM, simplificando o processo. O agente permite, de forma estruturada, integrar as tabelas e planilhas com as ferramentas do ecossistema Google, como Google Colab, Google Drive, Google Sheets e a API do Google Gemini. As classificações resultantes e códigos gerados pelos modelos foram armazenados em tabelas, gerando um novo conjunto de dados que possibilita a comparação entre os dados gerados pelos humanos e pela LLM.

A análise dos resultados foi composta por duas abordagens complementares. A primeira, uma análise quantitativa, contabilizando as convergências e divergências entre a codificação humana e a codificação realizada pelo LLM. A segunda abordagem consiste em uma análise qualitativa, por meio da revisão manual dos casos em que houver divergência entre o modelo e os códigos atribuídos pelos analistas humanos, com o objetivo de identificar possíveis padrões de erro ou limitações interpretativas do LLM. Por fim, os resultados obtidos serão discutidos com especialistas no tema, que participaram, inclusive, da codificação humana realizada para montar o conjunto de dados original.

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

Como resultado da análise das questões de pesquisa, foram geradas tabelas contendo informações quantitativas e qualitativas referentes ao conjunto de dados sobre Empatia na Engenharia de Software.

As tabelas geradas para suportar as análises incluem as seguintes colunas: Quotes (trechos de texto analisados), Category Manual (Baseline) (classificação atribuída manualmente pelos especialistas), Gemini 2.5 Pro (Prompt One Sort) (classificação atribuída pelo modelo LLM), Justifications Gemini 2.5 Pro (Prompt One Sort) (justificativas textuais fornecidas pelo modelo para a escolha da classificação), e Type of Divergence (tipo de divergência identificada entre a codificação manual e automatizada). Outras colunas complementares também foram incluídas para registrar aspectos relevantes à análise, como a complexidade do trecho ou observações dos pesquisadores.

Com base nas tabelas geradas, foi realizada uma análise quantitativa comparando as das classificações do Gemini 2.5 Pro, com o considerado padrão ouro, definido pelos pesquisadores humanos. Essa análise seguiu três critérios de equivalência: Equivalência

Total, Equivalência Parcial e Nenhuma Equivalência. A Equivalência Total foi considerada quando o código atribuído pelo LLM a um trecho textual coincidia exatamente com o código manual dos especialistas. A Equivalência Parcial foi aplicada nos casos em que o modelo atribui múltiplos códigos a um trecho. A categoria Nenhuma Equivalência indicou ausência completa de correspondência. serviu de base para a etapa subsequente de análise qualitativa dos casos divergentes. As Figuras 1 e 2 apresentam recortes demonstrativos dos dados quantitativos gerados.

Emotional Sharing	Embodiment, Understanding	The phrase "walking in each other's shoes" is a classic metaphor that directly represents **Embodiment**. The description of developing "a sense for how other people feel" and the explicit statement that it means to "understand" are clear indicators of the **Understanding** code, which involves grasping another's feelings and experiences.	Divergência Não Aceitável/Incorreta

Figura 1: Exemplo de caso em que houve discordância entre as classificações feitas por pesquisadores humanos e LLM.

Dados Incompletos	**0**	**0,00%**
TOTAL	**54**	**100,00%**
TOTAL PONTUAÇÃO	**40,48**	**74,96%**

Figura 2: Recorte de uma das tabelas contendo a taxa de concordância entre as classificações realizadas por pesquisadores humanos e LLM, para uma das questões de pesquisa abordadas.

Para além de todo o material gerado que serve como base para continuação do estudo (scripts, prompts, planilhas, tabelas e datasets), a análise qualitativa revelou insights relevantes que não poderiam ser captados por métricas quantitativas. Foi possível identificar padrões de interpretação adotados pelo LLM, além de limitações contextuais que influenciaram suas decisões. Em alguns casos, o modelo apresentou classificações alternativas plausíveis, que embora diferentes das originais, revelaram novas perspectivas interpretativas sobre os dados. Além disso, as justificativas geradas permitiram compreender melhor o raciocínio do modelo. Esses achados serão discutidos com especialistas no tema. A Figura 3 apresenta um recorte demonstrativo de um destes cenários.

O LLM tem mais Razão	O código do LLM, "Emotional Sharing", foca no tema central e explícito do trecho, que define repetidamente a empatia em termos de "vínculo emocional". Embora o código humano "Perspective Taking" possa ser inferido da menção à "similaridade de contextos", este é um mecanismo secundário para atingir o vínculo emocional, e não o foco principal da definição. A codificação do LLM é mais precisa por se concentrar no conceito dominante do texto.	O código do LLM é superior por sua centralidade e precisão em relação à definição explícita do trecho, enquanto o código adicional do humano, embora plausível, refere-se a um conceito secundário e menos enfatizado.

Figura 3: Recorde demonstrativo do cenário em que supõe-se que a classificação gerada pelo LLM é mais plausível que a classificação realizada por pesquisadores humanos.

CONSIDERAÇÕES FINAIS (ou Conclusão)

Os resultados até então obtidos neste experimento demonstram que modelos de linguagem de grande escala, como o Gemini 2.5 Pro, apresentam potencial significativo para apoiar a análise qualitativa de dados na área de fatores humanos em Engenharia de Software, especialmente em tarefas de classificação e codificação de trechos textuais. O estudo demonstrou a importância de desenvolver uma estrutura de agentes para propiciar um ambiente acessível, reprodutível e colaborativo para a condução dos testes e análises do experimento. Além disso, a construção dos prompts demonstrou ser um fator determinante na performance do modelo. O estudo demonstrou também que ainda há limitações importantes, especialmente em casos mais sutis ou contextualmente complexos, nos quais divergências aparecem de forma mais frequente. As justificativas fornecidas pelo modelo também contribuíram para a compreensão do “raciocínio” do modelo. Desta forma, conclui-se que, com a continuidade do estudo, futuras pesquisas poderão utilizar-se dos resultados e achados deste experimento para aprofundar o uso de técnicas de engenharia de prompt e personalização dos modelos para contextos específicos, com o objetivo de aprimorar ainda mais sua performance interpretativa, não só na Engenharia de Software. Para isso, o pacote experimental contendo scripts e as tabelas de resultados geradas está sendo disponibilizado (https://github.com/bielcmoraes/LLM-and-Software-Engineering).

REFERÊNCIAS

Bijker, R., Merkouris, S. S., Dowling, N. A., e Rodda, S. N. Chatgpt for automated qualitative research: content analysis. Journal of medical Internet research, 26:e59050, 2024.

CAPRETZ, Fernando L. Bringing the human factor to software engineering. *IEEE Software*, v. 31, n. 2, p. 104, 2014.

Cerqueira, L., Freire, S., Neves, D. F., Bastos, J. P. S., Santana, B., Spínola, R., Mendonça, M., e Santos, J. A. M. Empathy and its effects on software practitioners’ well being and mental health. IEEE Software, 2024.

SANTANA, B. et al. Psychological safety in the software work environment. *IEEE Software*, p. 1–8, 1 jan. 2024.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., e Oudeyer, P.-Y. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In Companion proceedings of the 28th international conference on intelligent user interfaces, p. 75–78, 2023.