

Universidade de Pernambuco Programa de Pós-Graduação em Engenharia da Computação (PPGEC)
Proposta de Dissertação de Doutorado
Área: Computação Inteligente
Título: “AgoraSim – Um Sandbox para Agentes de IA Conversacional”
Orientador: Pablo Barros <pablovin@gmail.com> Co-Orientadores: Pietro Gravino <pietro.gravino@sony.com>, Alessandra Sciutti <alessandra.sciutti@iit.com>
Descrição: A conversação é a alma da democracia [1], mas as plataformas online transformaram a escala e a natureza do diálogo, frequentemente promovendo polarização e toxicidade. Recentemente, agentes de IA têm mostrado potencial como intervenções para melhorar a saúde das conversas [2,3]. No entanto, existe uma lacuna crítica: a ausência de ambientes conversacionais realistas onde tais agentes possam ser projetados e testados de forma robusta antes de sua implantação no mundo real. Sem isso, corremos o risco de consequências não intencionais e de limitações na generalização dos resultados. Este projeto busca preencher uma lacuna fundamental na pesquisa em IA ao desenvolver um sandbox de conversação que simula discussões online realistas e em constante evolução. Fundamentado em dados reais de redes sociais e equipado com métricas de avaliação multidimensionais (como engajamento, toxicidade, concordância e deriva semântica), este ambiente utilizará o potencial de grandes modelos de linguagem (LLMs) para capturar toda a complexidade e nuances das conversas entre múltiplas pessoas. O projeto está estruturado em três fases interconectadas: Métricas de Avaliação de Conversa: Será definida e proposta uma suíte abrangente de métricas para quantificar a saúde de uma conversa, capturando não apenas toxicidade e polarização, mas também níveis de engajamento, padrões de concordância e discordância, e mudanças no conteúdo semântico ao longo do tempo. Desenvolvimento do Sandbox: Com base em dados do mundo real, será construído um simulador de conversas, treinado com tecnologia avançada de LLMs. Este sandbox irá reproduzir as dinâmicas das interações multi-usuários, permitindo que pesquisadores observem e experimentem em um ambiente controlado, mas realista. Testes e Avaliação de Agentes: Dentro desse ambiente simulado, diferentes agentes de IA conversacional poderão ser inseridos e testados rigorosamente. Permitindo que usuários humanos interajam com esses agentes no sandbox, a pesquisa irá criar cenários próximos do real que revelam como os agentes influenciam tanto a qualidade das conversas quanto o comportamento humano. Essa abordagem garante que os agentes sejam avaliados não apenas em simulações artificiais, mas também em contextos que espelham de perto as interações online reais. Ao entregar uma plataforma prática e extensível para o design e avaliação de IA agente em ambientes sociais complexos, este trabalho pretende lançar as bases para intervenções mais seguras e eficazes voltadas à melhoria da qualidade do discurso público. Este projeto será desenvolvido em parceria entre a Universidade de Pernambuco, o Sony Computer Science Labs (Paris, França) e o Istituto Italiano di Tecnologia (Genova, Itália). Uma sólida compreensão do idioma inglês é essencial para a sua realização.
Referências: [1] https://journals.sagepub.com/doi/10.1177/2057047316628310 [2] https://www.science.org/doi/10.1126/science.adq2852 [3] https://www.pnas.org/doi/abs/10.1073/pnas.2311627120