Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Doutorado

Área: Computação Inteligente

Título: Um Ambiente para Avaliação de Agentes Conversacionais em Ambientes Sociais Controlados

Orientador(a): Pablo Barros (pablovin@gmail.com)
Co-Orientador(a):
Pietro Gravino (pietro.gravino@sony.com) (Sony Computer Science Lab Paris - France)
Alessandra Sciutti (alessandra.sciutti@iit.com) (Istituto Italiano di Tecnologia - Italia)

Descrição:

As conversas online moldam opiniões, comportamentos e decisões coletivas, mas também estão cada vez mais associadas à polarização, conflitos e discursos tóxicos[1]. Embora agentes de inteligência artificial já demonstrem potencial para mediar, moderar ou melhorar interações sociais online [2], ainda falta um elemento fundamental: **ambientes experimentais realistas onde esses agentes possam ser projetados, treinados e avaliados antes de serem colocados em contato com pessoas reais [3]**.

Este projeto propõe a criação de um **sandbox de simulação de conversas online**, no qual discussões entre múltiplos participantes — humanos e agentes artificiais — possam ser reproduzidas, observadas e manipuladas de forma controlada. Utilizando **grandes modelos de linguagem (LLMs)** como base [4], o sandbox será capaz de simular dinâmicas reais de fóruns, redes sociais e discussões públicas, incluindo desacordos, formação de grupos, escalada de conflitos e mudanças de opinião ao longo do tempo.

O desafio central do projeto é responder à pergunta:
como projetar, avaliar e comparar agentes conversacionais que influenciam a qualidade de discussões sociais complexas?

Para isso, o projeto será estruturado em três componentes práticos e fortemente interligados.

No primeiro, será definida uma **suíte de métricas computacionais para avaliar a "saúde" de uma conversa**, indo além da simples detecção de toxicidade. As métricas incluirão engajamento, padrões de concordância e discordância, polarização, evolução semântica dos tópicos e estabilidade do diálogo ao longo do tempo. Essas métricas permitirão comparar diferentes estratégias de agentes de forma objetiva e reproduzível.

No segundo componente, será desenvolvido o **sandbox conversacional propriamente dito**: um simulador multiusuário baseado em dados reais de redes sociais e fóruns online, no qual LLMs representarão participantes com perfis, opiniões e comportamentos distintos. Esse ambiente permitirá a execução de experimentos controlados, nos quais o impacto de diferentes intervenções algorítmicas pode ser observado sem os riscos éticos associados a testes diretos em plataformas reais.

Por fim, o projeto permitirá a **inserção e avaliação de agentes de IA conversacional** dentro do sandbox. Esses agentes poderão atuar como moderadores, mediadores ou participantes ativos da conversa. Em uma etapa avançada, usuários humanos também poderão interagir com o ambiente, possibilitando experimentos híbridos humano–IA. O objetivo é analisar como diferentes tipos de agentes influenciam a dinâmica da conversa, o comportamento humano e os indicadores de qualidade do diálogo.

Ao final, o projeto entregará uma **plataforma experimental aberta e extensível**, capaz de servir como base para pesquisas futuras em IA social, moderação automática, agentes conversacionais e saúde do discurso público. O trabalho contribui diretamente para o desenvolvimento de intervenções mais seguras, transparentes e eficazes em ambientes sociais digitais complexos.

Este projeto será desenvolvido em colaboração internacional entre a Universidade de Pernambuco, o Sony Computer Science Labs (Paris, França) e o Istituto Italiano di Tecnologia (Gênova, Itália). É essencial que o(a) estudante tenha boa capacidade de leitura, escrita e comunicação em inglês.

Referências Bibliográficas:

[1] Shah, D. V. (2016). Conversation is the soul of democracy: Expression effects, communication mediation, and digital media. Communication and the Public, 1(1), 12–18.
[2] Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web (WWW '17), 1391–1399.
[3] Muñoz, P., et al. (2024). Quantifying polarization in online political discourse. EPJ Data Science.
[4] Argyle, L. P., Bail, C. A., et al. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. Proceedings of the National Academy of Sciences (PNAS), 120(30).