22ª Semana Nacional de ciência e tecnologia

Planeta Água: a cultura oceânica para enfrentar as mudanças climáticas no meu território

Desenvolvimento de *guardrails* para controle e marcação de mensagens na interação de usuários em sistemas baseados em grandes modelos de linguagem

Ricardo Augusto Franco | ricardoaugustofranco@hotmail.com
Eli Lopes da Silva | eli.lopes@ifsc.edu.br
Cristiano Mesquita Garcia | cristiano.garcia@ifsc.edu.br

RESUMO

Esta é uma pesquisa que faz parte de um trabalho de conclusão de curso do bacharelado em Sistemas de Informação do IFSC Câmpus Caçador. As soluções baseadas em inteligência artificial estão cada vez mais presentes, impactando diversos setores e gerando avanços significativos em várias áreas. A adoção crescente desses modelos requer a implementação de mecanismos que garantam a qualidade e a segurança das interações, especialmente em contextos sensíveis como o ambiente de cobrança, onde é essencial identificar conversas que não agregam valor ao negócio para evitar custos desnecessários para a empresa. Este trabalho tem como objetivo o desenvolvimento de guardrails para controle e marcação de mensagens enviadas por usuários durante interações em sistemas de cobrança baseados em grandes modelos de linguagem, os Large Language Models (LLMs). Para isso, propõe-se a criação de uma Interface de Programação de Aplicações (API), que utiliza técnicas de processamento de linguagem natural para analisar as mensagens. A API classifica e marca interações que apresentem padrões de comportamento prejudiciais, permitindo respostas rápidas das equipes de cobrança e a criação de datasets mais precisos. A implementação dos guardrails, ao focar nas mensagens dos usuários, busca filtrar interações que possam resultar em ações indevidas, como fraudes ou disputas maliciosas.

INTRODUÇÃO

Empresas brasileiras adotam cada vez mais a comunicação via mensagens para cobrança e relacionamento com clientes. No entanto, interações mal-intencionadas consomem recursos de IA sem retorno ao negócio. Soluções com grandes modelos de linguagem (LLMs) exigem mecanismos de segurança que mantenham a eficiência e a integridade dessas interações.

OBJETIVO

Desenvolver *guardrails* capazes de identificar, classificar e marcar mensagens que não geram valor, reduzindo custos operacionais e apoiando decisões rápidas no setor de cobrança.

METODOLOGIA

✅ Pesquisa exploratória com revisão bibliográfica e análise de casos.
✅ Aplicação do modelo CRISP-DM.
✅ Seleção de técnicas de NLP e *autoencoders* para detecção de anomalias.
✅ Comparação de padrões entre mensagens legítimas e prejudiciais.
✅ Desenvolvimento de API para integração com canais digitais de cobrança .

RESULTADOS PARCIAIS

A classificação é complexa devido ao uso diverso e coloquial da língua portuguesa.
Dados altamente desbalanceados (menos de 10% são interações sem valor).
Necessidade de técnicas específicas de balanceamento e mitigação de vieses.
Indicadores iniciais mostram potencial para reduzir custos e melhorar a precisão operacional .

CONCLUSÕES

O uso de *guardrails* em sistemas com LLMs promove:
✅ Segurança e responsabilidade no uso da IA.
✅ Conformidade com LGPD.
✅ Maior eficiência nos fluxos de cobrança.
✅ Sustentabilidade financeira na operação de modelos generativos.
✅ Fortalecimento do relacionamento com clientes reais.

REFERÊNCIAS

CHAPMAN, P. *et al*. CRISP-DM 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, n. 13, p. 1–73, 2000.

FRENAY, L. et al. **Messaging as a Value Driver for Brazilian Businesses**. 2024. Disponível em: https://web-assets.bcg.com/b9/32/2fdce0d9409780446961596f5aee/messaging-as-a-value-driver-for-brazilian-businesses.pdf. Acesso em: 7 dez. 2024.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo, Brasil: Atlas, 2002.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016.

GUANAES, Lucas. O discreto monopólio do meta: um estudo de caso sobre o domínio do Whatsapp no Brasil. **Seminário do LEG**, Limeira, SP, v. 14, n. 1, p. 29–41, 2024.

LIMNA, P. *et al*. The use of chatgpt in the digital era: perspectives on chatbot implementation. **Journal of Applied Learning and Teaching**, v. 6, n. 1, p. 64–74, 2023.

MARCONI, M. A.; LAKATOS, E. M. **Metodologia científica**. 7. ed. São Paulo: Atlas, 2017.

VASWANI, A. *et al*. Attention Is All You Need. **arXiv preprint arXiv:1706.03762**. 2023. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 15 out. 2025.

CNPq
INSTITUTO FEDERAL Santa Catarina
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO